A FRAMEWORK FOR ENHANCING SPEAKER AGE AND GENDER CLASSIFICATION BY USING A NEW FEATURE SET AND DEEP NEURAL NETWORK ARCHITECTURES

Arafat Abu Mallouh

Under the Supervision of Dr. Buket D. Barkana

DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

AND ENGINEERING

THE SCHOOL OF ENGINEERING

UNIVERSITY OF BRIDGEPORT

CONNECTICUT

November, 2017

A FRAMEWORK FOR ENHANCING SPEAKER AGE AND GENDER CLASSIFICATION BY USING A NEW FEATURE SET AND DEEP NEURAL NETWORK ARCHITECTURES

Arafat Abu Mallouh

Under the Supervision of Dr. Buket D. Barkana

Approvals

Committee Members

Name	Signature	Date
Dr. Buket D. Barkana	Buket Barkana	11/3/2017
Dr. Xingguo Xiong	Xingguo Xiong	11/03/2017
Dr. Navarun Gupta		11/3/17
Dr. Miad Faezipour	Miad Faezipour	11, 3, 2017
Dr. Amalia Rusu	Amalia Rusu	11/3/17

Ph.D. Program Coordinator

Dr. Khaled M. Elleithy		12/13/17

Chairman, Computer Science and Engineering Department

Dr. Ausif Mahmood	Ausif Mahmood	12-13-2017

Dean, School of Engineering

Dr. Tarek M. Sobh		12-14-2017

A FRAMEWORK FOR ENHANCING SPEAKER AGE AND GENDER CLASSIFICATION BY USING A NEW FEATURE SET AND DEEP NEURAL NETWORK ARCHITECTURES

A FRAMEWORK FOR ENHANCING SPEAKER AGE AND GENDER CLASSIFICATION BY USING A NEW FEATURE SET AND DEEP NEURAL NETWORK ARCHITECTURES

ABSTRACT

Speaker age and gender classification is one of the most challenging problems in speech processing. Recently with developing technologies, identifying a speaker age and gender has become a necessity for speaker verification and identification systems such as identifying suspects in criminal cases, improving human-machine interaction, and adapting music for awaiting people queue. Although many studies have been carried out focusing on feature extraction and classifier design for improvement, classification accuracies are still not satisfactory. The key issue in identifying speaker's age and gender is to generate robust features and to design an in-depth classifier. Age and gender information is concealed in speaker's speech, which is liable for many factors such as, background noise, speech contents, and phonetic divergences.

In this work, different methods are proposed to enhance the speaker age and gender classification based on the deep neural networks (DNNs) as a feature extractor and classifier. First, a model for generating new features from a DNN is proposed. The proposed method uses the Hidden Markov Model toolkit (HTK) tool to find tied-state triphones for all utterances, which are used as labels for the output layer in the DNN. The

DNN with a bottleneck layer is trained in an unsupervised manner for calculating the initial weights between layers, then it is trained and tuned in a supervised manner to generate transformed mel-frequency cepstral coefficients (T-MFCCs). Second, the shared class labels method is introduced among misclassified classes to regularize the weights in DNN. Third, DNN-based speakers models using the SDC feature set is proposed. The speakers-aware model can capture the characteristics of the speaker age and gender more effectively than a model that represents a group of speakers. In addition, AGender-Tune system is proposed to classify the speaker age and gender by jointly fine-tuning two DNN models; the first model is pre-trained to classify the speaker age, and second model is pre-trained to classify the speaker gender. Moreover, the new T-MFCCs feature set is used as the input of a fusion model of two systems. The first system is the DNN-based class model and the second system is the DNN-based speaker model. Utilizing the T-MFCCs as input and fusing the final score with the score of a DNN-based class model enhanced the classification accuracies. Finally, the DNN-based speaker models are embedded into an AGender-Tune system to exploit the advantages of each method for a better speaker age and gender classification.

The experimental results on a public challenging database showed the effectiveness of the proposed methods for enhancing the speaker age and gender classification and achieved the state of the art on this database.

ACKNOWLEDGEMENTS

Completion of my doctoral dissertation was possible with the support of several people. I would like to express my sincere gratitude to all of them. I would like to express my special appreciation and thanks to my advisor Dr. Buket D. Barkana, she has been a tremendous mentor for me. I would like to thank her for encouraging my research and for allowing me to grow as a research scientist.

I would also like to thank my committee members, Dr. Navarun Gupta, Dr. Xingguo Xiong, Dr. Miad Faezipour, and Dr. Amalia Rusu for serving as my committee members even at hardship. I also want to thank you for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you. I am grateful for Dr. Khaled Elleithy, the director of the Ph.D. program, for the academic support and the facilities provided to carry out the research work at the Institute.

Nobody has been more important to me in the pursuit of my Ph.D. degree than the members of my family. I owe everything good in my life to my mother and father for all of the sacrifices that they have made on my behalf. Your prayer for me was what sustained me thus far, your love and guidance are with me in whatever I pursue. Most importantly, I wish to thank my beloved brother, Jamal, who has provided unending inspiration and he has supported me more than I could ever give him credit for here.

I would like to thank all the staff and colleagues of the School of Engineering for their support that made my study in the University of Bridgeport a wonderful and exciting experience.

A great warm word for my dear friend and fellow researcher Zakariya Qawaqneh, Zakariya's continuous support, deep discussions, and persuasive analytical thinking have enriched our research and always pushed our goals to maximum limits. Zakariya was always there during the happy and hard times. For all these reasons and many, many more, I am eternally grateful. Thank you Zakariya for making the long journey of pursuing our Ph.D. degrees bearable.

Above all, I owe it all to Almighty God for granting me the wisdom, health and strength to undertake this research task and enabling me to its completion.

TABLE OF CONTENTS

LIST OF TABELS

LIST OF FIGURES

ABBREVIATIONS

aGender	Age-Annotated Database of German Telephone Speech
ASR	Automatic Speech Recognition
AUC	Area Under Curve
BN	Bottleneck
CMVN	Cepstral Mean Variance Normalization
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DNNs	Deep Neural Networks
DT	Decision Tree
GRNNs	General Regression Neural Networks
HMMs	Hidden Markov Models
HRI	Human-Robot Interaction
HTK	Markov Model Toolkit
KNN	K-Nearest Neighbors

LDA	Linear Discriminant Analysis
LPA	Linear Prediction Analysis
LPCC	Linear Prediction Coding Coefficients
MCM	MFCCs-Class Models
MCMS	Mel Cepstral Modulation Spectrum
MLLR	Maximum Likelihood Linear Regression
MLP	Multi-Layer Perceptron Network
MSM	MFCCs-Speakers Models
NB	Naïve Bayes
PLP	Perceptual Linear Prediction
PPR	Parallel Phone Recognizer
RBM	Restricted Boltzmann Machine
ROC	Receiver Operating Characteristics
SCM	SDC Class Models
SDS	Spoken Dialogue Systems
SSM	SDC Speaker Models
T-MFCCs	Transformed Mel-Frequency Cepstral Coefficients
TPP	Tandem Posteriors Probability

UF-VAD	University of Florida Vocal Aging Database
VAD	Voice Activity Detection
WSNMF	Weighted Supervised Non-Negative Matrix Factorization

CHAPTER 1: INTRODUCTION

Age and gender classification is defined as the extraction of age and gender information from speaker's speech. A key stage in identifying speakers' age and gender is to extract and select effective features that represent the speaker's characteristics uniquely. Another key stage is classifier design. A classifier uses the extracted features to predict the speakers' age and gender. The focus of this research is on finding distinctive feature set that is able to represent the speaker identity such that the classifier can recognize the age and gender of the speaker efficiently. In addition, the design of the classifier plays a major role in classifying the speaker's age and gender. This research investigates different classifiers and classification techniques to enhance age and gender classification.

Numerous feature sets have been developed and evaluated in the literature for this problem. Those features can be classified into three categories, spectral, prosodic, and glottal features. One of the most recognized feature sets is MFCCs which represent the spectral characteristics of speech utterance. MFFCs are widely used in the literature for different speech processing applications such as speech recognition, speaker identification, and noise classification. MFCCs represent the spectrum that is related to vocal tract shape and do not capture the prosodic information [1]. The effectiveness of MFCCs comes from the ability to model the vocal tract in short-time power spectrum. There are studies reporting high overall classification accuracies [2] (around 90%), however these studies either used a small private corpus or predicted a small number of age and gender classes.

Several classifiers have been used in the literature for speaker's age and gender classification with different levels of performance. One of the most recent popular techniques is the eigenvoice (I-Vector) which is based on the process of joint factor analysis [3]. Currently, it is considered as one of the state-of-art in the field of speaker recognition and language detection [4, 5]. Eigenvoice adaptation is the main procedure to estimate I-Vector which represents a low-dimensional latent factor for each class in a corpus. A test data is scored by a linear strategy that computes the log-likelihood ratio between different classes. Another popular classifier that used in speech field is the GMM, GMM is considered to be one of the most effective models that have been used in different fields such as speaker recognition, language identification, and speaker age and gender classification. It is used as a standard classifier for text-independent speaker recognition because of its ability to approximate various arbitrary shaped distributions. One of the most attractive benefits of GMM model is its fast training process compared to other models [6].

1.1 Main Speech-Related Fields

Speech processing is one of the main fields that gives theoretical and practical insight for different methodologies of how the machines can deal with the speech signals (human speech). Certainly, as it is known, human speech contains various information such as speaker specific characteristics, emotions, and language context. Therefore, processing and extracting the basic parameters of the speech signal are needed to capture this information. Recently, there have been significant advances in different speech-related fields. The speech signal analysis has shown great promise in various life science applications. Examples of the main related speech fields are automatic speech recognition, speaker recognition, language recognition, speaker

accent identification, speaker emotion recognition, and speaker age and gender classification.

1.1.1 Automatic speech recognition

ASR is one of the active research topics that tries to teach an independent machine-based the ability to identify and process human speech. By identifying the speech, the machine can use the decoded speech as input for a wide-range of real applications. For example, call steering, identification for security usage, handling customers enquiries, and for computer dictation. The speech signal carries linguistic information and speaker dependent information such as the age, gender, emotional state, and some ethnic features. There are many factors that affect the robustness of any ASR system, for example, the spectral density of the speech, speech segments, context dependent, accent, and pronunciation. Developing a robust ASR system requires a set of reliable techniques which plays a major role for performing a successful speech recognition, for example, efficient feature extraction techniques to capture speech and speaker variability, acoustic modeling techniques, pronunciation modeling techniques, and diverse training benchmarks. ASR has been studied earlier in the literature as in [7-10], and recently, major research efforts have been focused in order to enhance ASR by using new methods and pioneer ideas as in [11-16].

1.1.2 Language recognition

Language Recognition can be defined as the ability to identify automatically the spoken language using machine-based solutions. This field has different applications in real life and its importance is expected to increase in the near future due to rapid

development of communication technology in the world [17-18]. For example, automatic language recognition is used in emergency call forwarding, customer service centers, and translation systems deployed in multilingual environments such as international conferences and huge medical centers. Normally, language recognition depends on spoken sounds or on the pronunciation dictionary together with a transcription of the spoken text. Different research has been carried out in the past and recently for making automatic language recognition possible and accurate such as [19-24]. There are a set of keys that help machine and human to differentiate between languages, generally, languages differ in many aspects such as:

- prosody: stress, duration, and pitch.
- syntax: the patterns and structure of the sentences.
- Morphology: roots and lexicons.
- Phonology: some phonemes are different between languages.

1.1.3 Accent recognition

Speaker accent is one of the speaker variability factors that makes speech recognition a challenging task. Accent recognition can be defined as the ability to recognize the accent of the speaker automatically within a predetermined language using the speaker voice [25-26]. This type of recognition contributes in other recognition tasks such as language identification and speech recognition. Accent recognition can be applied for foreign accent or for domestic accent recognition. Different methods and techniques have been proposed in the literature for making the accent recognition process more reliable and to increase the accuracy rate as in [27-34].

1.1.4 Speaker emotion recognition

The interaction between the human and computer is increasing rapidly and offers various kinds of services. One of the factors that have a major role for enhancing the interactions is the ability to recognize the emotion state related with the recognized speech. If the human orders the computer to do perform some action then the computer will respond depending on its ability to recognize the spoken words. On the other hand, if the computer can identify the emotion that is associated with recognized speech then it can respond in different ways depending on the emotional state of the speaker and the wanted service. Different studies have been performed to understand what features that affects the recognition of the speaker emotions as in [35-37], and other studies focused on finding solutions to enhance the accuracy of the recognition rates as in [38-41].

1.1.5 Speaker recognition

Speaker recognition is one of the most popular field in speech research over the last decades. This field consists of two main subfields, speaker identification and speaker verification. On the last decades, several studies have been conducted on speaker recognition using different techniques. These techniques can be summarized and categorized into four categories: Vector quantization based systems [42-49], GMM based system [50-55], factor analysis based system [56-59], and most recently deep learning bases systems [60-66].

- Speaker identification: is the process of determining whose speaker provides a given speech. In the speaker identification process the number of decision depends on the number of population in the used databases, therefore the system

performance of speaker identification will decrease if the size of the speakers used to build the system increases. In general, in any speaker identification system, a given speech utterance (speech signal) is processed and analyzed to be compared with different models for known speakers. Then the given speech (the unknown speaker) is identified as the speaker who best matches the known identified models.

- speaker verification: is the process of verifying the identity of the speaker based on his/her speech. In simple words in this field, a given speech utterance for unknown speaker is compared with the speaker model whose identity is being claimed. If it passes a threshold, then the claimed identity is verified and accepted otherwise the identity is rejected. Choosing an optimal threshold for accepting and rejecting the claimed identity is one of the most critical issue for speaker verification. Choosing a high threshold leads to prevent most of unauthenticated users (imposters) to get access for the system, but this also will increase the risk of rejection the authenticated users to get access to the system. on the other hand, choosing a low threshold increase the risk of accepting the unauthenticated users even it will give an access for the authenticated users in most cases. Therefore, choosing and optimal threshold should be taken in account based on the distribution of the unauthenticated users and the authenticated users in the new system.

1.2 Motivation Behind the Research

Currently, computerized systems such as language learning, phone ads, criminal cases, computerized health and educational systems are rapidly spreading and imposing an urgent need for better performance. Such applications can be improved by speakers' age, gender, accent, and emotional state information [67-69]. Moreover, many of the daily life

activities associated with humans' life style are being computerized, some of these activities are related either to the health of the human or to services needed to perform essential daily tasks. These activities rely on visual or speech data input where the ability to efficiently recognize this input controls the quality of the provided service. Recently, remarkable advancement in hardware and software tools creates new opportunities and open new doors for solving and improving different research problems. For example, DNNs have been used effectively for feature extraction and classification in computer vision [70-71], image processing and classification [70, 72], and natural language recognition [73- 74]. In 2006, Hinton et al. [75] introduced the restricted Boltzmann machine (RBM) for the first time as a keystone for training deep belief network (DBN). Later, Benjio [76] successfully proposed a new way to train DNN by using auto encoders. DNN has a deep architecture that transforms rich input features into strong internal representation [77]. Although previous studies have presented some improvements in speaker age and gender classification, the classification of speaker's age and gender still has a big room for improvement.

More effective feature sets, especially for short-time duration speech utterances, and classifier designs are required to improve current classification accuracies. In addition, the architecture of the classifier could be improved to enhance the accuracy of the task.

1.3 Main Contributions of the Proposed Research

Improving speaker's age and gender classification is achieved through incorporating a set of related steps that together can enhance the classification accuracy. The first step is to extract a new feature set from speech utterances that contains rich

information about the speaker. On the same time, the feature set should be compact and include a high representation of the input data in order to permit the classifier to find unique patterns for different speakers easily. The second step is to introduce new classification architectures and techniques that can benefit from the distinguished nature of the DNNs as feature extractors and classifiers. Finally, the cooperation of the new feature set and classification architectures should be optimized to utilize their capabilities together.

In this research, MFCCs features are investigated to produce a new feature set that contains more information about the age and gender of the speaker. As well as, a new classification techniques will be introduced for enhancing the classification. In this research, the key contributions are listed as follows:

- A new feature set which is called T-MFCCS is introduced in order to capture more distinctive information about the age and gender from speaker speech. The new feature set is developed using the tied-state triphones. The tied-state triphones is extracted using the HTK tool for all training and test utterances, then they are used as labels for the output layer of a DNN classifier with a bottleneck layer. After training the DNN, the output of the bottleneck layer is used to generate the new T-MFCCs features.
- DNN-based speaker models using the SDC feature set are proposed in order to improve the classification accuracies in speaker age and gender classification. A model for each speaker is built instead of using one model for each class of speakers. Introducing a speaker-aware model is motivated by the fact that a speaker model can capture the characteristics of the speaker more effectively than a model that represents a group of speakers.

- AGender-Tune system is proposed to classify the speaker age and gender by jointly fine-tuning two DNN architectures; Age DNN to classify speaker age, and Gender DNN to classify the gender. A third output layer is proposed to combine the output layers of Age and Gender DNNs using element-wise summation.

CHAPTER 2: LITERATURE SURVEY

The problem of age and gender classification was studied early in 1950's [78], but the computer-aided systems for deriving the age and gender information from speech have been developed recently [79-80].

In [81] they developed a new acoustic feature set for estimating speaker's age. Their baseline feature set is the MFCCs which are extended by a set of prosodic features, pitch f0, and first four formant frequencies. The combination of these features results in 220 features, these 220 features are reduced by selecting the best feature subset by maximizing the R2 variance with R as correlation by using multiple regression/correlation analysis. Their approach selects the best subset that is composed of one feature, two features, and continues until there is no better subset. They tested their work on the University of Florida Vocal Aging Database (UF-VAD) which contains 5 hours of speech for 150 different speakers and 1350 utterances of read English speech. The UF-VAD has 3 age groups evenly divided between males and females for young, middle-aged, and old age groups. For each speaker in the database they generate a constant high-dimensional feature vector that is independent of the length of the length of the utterance and of the extracted features and is represented by a Gaussian model. They reported that adding prosodic, pitch, and formant features to the MFCCs feature set improved the results by reducing the mean absolute error between 4-20%.

[82] proposed to use the modulation cepstrum coefficients for classification age and gender instead of using the cepstral coefficients features. They extracted a smooth

information of the cepstral over a period of times (frames). for extracting frames from the speech utterance, the discrete cosine transform (DCT) is used over a fixed duration window. in other words, they filter the speech utterance in modulation cepstrum domain by decomposing the utterance cepstral trajectories into group of low and slow frequencies and extract the mel cepstral modulation spectrum (MCMS) features. They reported that the low modulation frequencies of MCMS (3-14 hertz) has the efficient information that need for the age and gender classification. They evaluated the efficiency of the proposed features set a total of around 6000 utterances collected from German SpeechDat-II corpus and VoiceClassData on different 7 age and gender classes. They compared the performance of the extracted MCMS features with the MFCCs. They reported an accuracy of 50.2% using the MCMS features.

[83] proposed and compared three novel systems which combines short-term cepstral features and long-term features for speaker age recognition. In their work they stated that acoustic analysis indicates that some specific features such as pitch extracted from span of speech correlate clearly with the speaker age despite the fact that common successful systems in the literature such as GMM models and multiple phone recognizers that utilize such features have less performance than other features such as GMM models with short-term cepstral features and multiple phone recognizers trained with the data of speakers of the respective class. The first system is SVMs using phone conditioned MFCCs plus utterance based pitch on the feature level, the feature set for this system is a set of segment-based features combined with utterance-based features. This feature set was used to train seven binary SVMs where each SVM represents a class and the model with highest likelihood is chosen for the decision. The second system is the SVMs using phone-

conditioned MFCCs plus utterance-based pitch combined on the score level. The Third system is the GMM models using frame-based MFCCs plus SVMs using utterance-based pitch and they combined their score-level results. For testing, the proposed system was evaluated on the German SpeechDat-II corpus, the database has 4000 native German speakers where 80 speakers of each age and gender group were selected for training and 20 for testing. The German SpeechDat-II have 7 age and gender groups. Also, they used a reference system that utilized jitter, shimmer, the mean and the standard deviation of the additive noise in the voice signal, and statistical derivatives of the F0 pitch including mean, standard deviation, and slop as features, resulted in 17-dimensional feature vector calculated for each sample. The reference system is the multi-layer perceptron network (MLP) which has one hidden layer and sigmoid used as an activation function. Three MLPs were used and trained on three feature sets to classify gender, female speakers age class, and male speakers age class. The average accuracy for the first system is 39.9%, the second system is 43.7%, and for the third system is 49.11%.

In [84] a system for detecting the older people over the spoken dialogue systems(SDS) to meet their needs is proposed. Authors in this paper try to distinguish elderly peoples from other speakers using two characteristics, interaction style and information processing speed(DSST). Several acoustic and lexical features were used which were extracted from the speech utterance such as pitch features, speaking rate, MFCCs, vocal tract length, and words frequency in the utterance. The DSST variable is calculated using Wechsler Adult Intelligence Scale subtest by dividing the digit over the symbol substitution. They found that the DSST variable average for old people is 51. While for the young people had average of 75. For interaction style, they found that there is big

differences between young and old people. Three feature sets are used to simulate the interaction style of the speaker,1) overall dialogue statistics 2) speech act group frequency 3) word group frequency. Each speaker has been clustered in each feature and in a combination of the all the features. They found that 62% of the old people used a social interaction style, while 4.2% of young people used a social interaction style. They used their own small collected data to evaluate the predicted features for predicting the old people.

In [85] proposed a system which combined five methods at the acoustic level for speaker age and gender identification: The systems were GMM system based on MFCCs features, SVM based on GMM mean supervector, GMM based on GMM maximum likelihood linear regression (MLLR) supervector, SVM based on GMM Tandem Posteriors Probability (TPP) supervectors, and SVM baseline system based on the 450-dimensional feature vectors which includes prosodic features at the utterance level. In addition, they combined two or more systems by using score level fusion technique to increase the classification accuracy. For the GMM system they used 13 MFCCs together with their first and second derivatives as a feature set, in total 39 features for each frame were extracted. Also, for zero mean and unit variance normalization they applied cepstral mean subtraction and variance normalization. They utilized a UBM in conjunction with MAP model adaptation technique because the training data for age and gender classes is small to train an efficient GMM. For the GMM-SVM mean supervector system, the means of Gaussian components were adapted using MAP for each UBM, training, and evaluation sets. To create the GMM supervectors they concatenated the mean supervectors of all Gaussian components which were modeled by SVM. To reduce the computation cost, they added

randomly one dummy dimension at the head of each mean supervector with value 1 so that the target score can be calculated using a inner product, also they used two stage classification framework. Finally, they performed mapping from the supervectors inti discriminative database characterization score vectors, For the GMM-SVM MLLR supervector system, they performed MLLR adaptation for each sample on the training and evaluation sets on the UBM. They used the MLLR matrix of supervectors for SVM modeling. Finally, they applied linear discriminant analysis (LDA) to reduce the dimension of the MLLR supervector. For the GMM-SVM TPP supervector system, they extracted the TPP features for each utterance in the training and evaluation sets on the UBM. They reported that larger posterior probability enables the Gaussian component to represent the feature vector. They showed that TPP supervectors contain age and gender specific information. In addition, the authors stated that the SVM system combined with MFCCs feature based system improve the classification performance since SVM system includes different prosodic features. Finally, the evaluation of their work was performed on the aGender database. The overall accuracies of the five individual classifiers were 43.1%, 42.6%, 36.2%, 37.8%, and 44.6%, respectively. The combined GMM and GMM-SVM mean supervector systems achieved 45.2% of overall accuracy. The fused classifier, the combination of GMM-SVM MLLR supervector and GMM-SVM TPP supervector systems, achieved an overall accuracy of 40.3%. The fusion of the first four classifiers achieved an overall accuracy of 50.4%. Finally, the fusion of the five classifiers performed slightly better by achieving overall accuracy as 52.7%.

[86] studied four approaches for automatic speaker age and gender classification based on telephone applications. Also, they compared the classification results with human

performance on the same data. The four automatic approaches were based on, a parallel phone recognizer (PPR); dynamic Bayesian networks to combine prosodic features; linear prediction analysis (LPA); and GMM based on the MFCCs feature set were compared. The first system based on the PPR is derived from automatic speech recognition and automatic language identification. The feature extraction process consisted of two parts, finding the MFCCs, and linear transformation, where 24 components were retrained for the feature vector. Also, they used specific phoneme recognizer with category specific HMM and bi-gram for each class. The second system use four feature sets, the first two features were prosodic (jitter and shimmer) features. the third feature set is the harmonics-to-noise-ratio which was calculated using the mean and the standard deviation of the utterance. The fourth feature calculated statistical features (F0, standard deviation, mean average slope) in total 17 features were calculated. They used two layers of classification, the first layer is three MLPs, and the second layer used Dynamic Bayesian Networks to model the classification-inherent uncertainty. The third system utilized the dependency between age and gender on the linear prediction envelope over a windowed speech signal. The distance between the signal spectrum and the linear prediction spectrum was measured. Then the GMM is used to estimate these distances for all training data. The fourth system utilized two independent frame-wise classifiers, then their decisions are combined at the utterance level. The first classifier was 256 independent GMM per class trained for the MFCCs and their first and second derivatives for age classification. The second classifier was trained to predict the gender of each class separately, a GMM of 128 mixtures were used over the MFCCs, pitch, and additional features. For evaluating their work, they collected their own database for training which consists of 7 age and gender groups. The SpeechDat-II corpus was used for

evaluation. Overall accuracies were reported as 54%, 40%, 27%, and 42%, respectively. Overall classification accuracy of human listeners was reported as 54.7%.

In [87] they investigate several systems for age and gender classification for human-robot interaction (HRI). Two different feature sets were used namely MFFCs and linear prediction coding coefficients (LPCC). As well as, two different classifiers were used, SVM and C4.5 decision tree(DT). In total, four combinations are compered 1) MFCC with SVM 2) MFCC with DT 3) LPCC with SVM 4) LPCC with DT. For age, they classified two groups children and adults. They collected their own database to evaluate the performance. The database for gender consists of 6960 utterances for training and 3482 utterances for testing. These utterances were collected from 7 females and 7 males. While for age, 12925 utterances were used for training and 6464 utterances were used for testing. The overall accuracies were reported as 91.39% and 88.37%, 84.69%, 82.72, respectively, by using MFCCs-SVM, MFCCs-DT, LPCC-SVM, and LPCC-DT for age classification. The overall accuracies for gender classification by using the same systems were calculated as 93.16%, 91.45%, 86.60, and 83.02 by the same classifiers. According to their results, the MFCC feature set with the SVM as classifier gave the best result for age and gender classification for the HRI system.

[88] built, compared, and combined 5 different systems to classify the age and gender of the speaker. The first, second and third systems are called GMM-Base group, since all the systems are model the features into supervectors that have concatenated mean vectors of GMM. As well as, these systems using three different feature sets, MFCCs, PLP, and TRAPS respectively. The fourth system which also called glottal system, from the voiced speech segment predict nine glottal features using two mass vocal model. Then 27

feature vectors for each one of the nine glottal features were extracted by calculating the minimum, maximum and the mean. Tow physical mass vocal model are extracted to be used in the fourth system and they are estimated in data driven 9 glottal features. The fifth system which called also prosodic system, extracted 219 prosodic features from each utterance for the voice and unvoiced speech segments. Then all the systems are fused in two ways to compare different combination of the systems, early fusion and late fusion. The early fusion is performed in the feature level. That means all the extracted features of the five systems were concatenated together to form a high dimension features size vector of size (3878 features). Then it is classified using the SVM. The late fusion system combines the five systems on their level score using a multi-class logistic regression. The effectiveness of the purposed systems was conducted using the aGender database. on the development set of aGender database the early fusion system achieved 46.1% accuracy, while the late fusion system achieved accuracy of 47.8%.

[89] studied fusion technique by using various individual classification systems. They used three classifiers, the GMM-UBM, MLP, and SVM, also the short and long-term acoustic and prosodic were used as features. They introduced the age system which consists of multiple age detection classifiers which have different features extracted from different training sets. The used classifiers were the SVM and the MLP. The features used were 12^{th} order Perceptual Linear Prediction (PLP) coefficients, energy, deltas, pitch (f0), 28 static and modulation spectrogram features. The output scores of this multi front-end system are calibrated and combined to find the final output. The age labels are extracted after the classification of the input data into seven age and gender labels. Moreover, they introduced the gender system which have the same architecture for the age system where multi training

sets, multi features, and multi classifiers were used. In their work, they compute three gender labels as male, female, and child after finding the probability score of the seven age and gender classes. For training and testing their work, they used four different corpora, aGender, CMU Kids corpus, PF STAR children corpus, and the BN ALERT corpus. The highest classification accuracy for age using the development set of aGender was 51.2%, and the highest accuracy for gender was 83.1%.

In [2] a system is built to be used in a home-robot for classification human age and gender from the speech. In their research, they focused to analysis the voice to obtain the information related to the human age and gender using the MFCCs. They noted that the used the MFCCs feature set since it is the well-known feature used in other speech fields. Moreover, MFCCS is more robust in the noisy environments and is not dedicated for vowels likes the other features used for age and gender classification such as jitter, shimmer, F0, and harmonics-to-noise ratio. The GMM is utilized to be used as a classifier. They formulated three different tests scenarios: 1) age test which consists of 2 classes (adult and child), 2) gender test which consists of 2 classes (male and female), 3) age-gender test which consists of 4 classes (male adult, female adult, and child). They tested the effectiveness of their system using their own data. They collected a database which is called ETRI-VoiceDB2006 under robot environment conditions that take in account the gap between the robot and the user. The adults age range is (22-45), while the children age range is (10-11). For age, they reported accuracy of 96.57%, while for the gender the accuracy was 94.9%.

Weighted supervised non-negative matrix factorization (WSNMF) and general regression neural networks (GRNNs) were used by Bahari et al. [90] to design an age and

gender regression system. They achieved an accuracy of 96% for gender recognition on a Dutch speech database. For age estimation, the achieved mean absolute error was 7.48 years.

Nisimura and Lee [91] proposed a speech guidance system and used an SVM classifier, which was able to classify adult and children speakers from a private database. Classification accuracy was reported as 92.4% by using acoustic and linguistic features of speech utterances. Dobry et al. [92] proposed a speech dimension reduction method for age-group classification and precise age estimation. After deploying SVM with RBF kernel, they noted that the classifier performance was improved by using their dimension reduction method and the SVM classifier was faster and less affected by over-fitting problem. In [80], Muller et al. built a special system for elderly people. Four classes as elderly female, non-elderly female, elderly male and non-elderly male were studied. Two databases were used, which were M31 and ScanSoft, to evaluate the system performance. Jitter, shimmer, and speech rate were used as features for k-nearest neighbors (KNN), SVM, and naïve Bayes (NB) classifiers.

CHAPTER 3: PROPOSED FEATURE SET AND METHODS BASED ON DNN

In this section, three different methods for improving speaker's age and gender classification are proposed. Each of the proposed methods focuses on one specific area for improving the problem. The first area is the feature set, the second area is the classification method, and the third area is the classifier architecture. Each method will be explained in detail and will be accompanied with any limitation or restriction for application. The rest of this chapter is organized as follows: Section 5.1 introduces a new feature set, Section 5.2 describes a new classification method, and finally section 5.3 presents new architectures for age and gender classification.

3.1 Transformed MFCCs Feature set for Speaker Age and Gender Classification

In this section, the generation of transformed features and the suggested regularized DNN weights using shared class labels are explained. An approach to transform existing features into more effective features is proposed. MFCCs, their first and second derivatives are used as input features for comparison reasons since most of the previous studies have used MFCCs features in age and gender classification [85-86].

3.1.1 Generation of transformed features

New transformed features are generated from input features by using DNN as shown in Figure 1. For example, glottal and spectral features can be used to generate a new form of features in speech field.

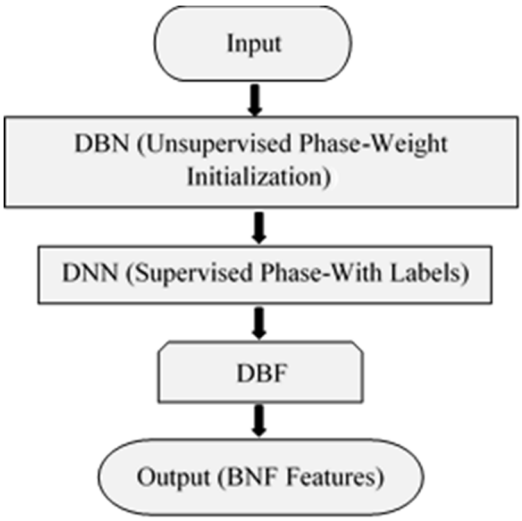


Figure 1. The main steps for extracting the BNF features from the input features.

The DNN that is used to generate these features consists of several hidden layers in which one of them has a very small number of units compared to other layers. The resulted features can be considered as a low-dimensional representation since the bottleneck layer compresses the input features and the output labels to form new features. It is as a way of nonlinear dimensionality reduction since it produces a low-dimensional feature set from the input features based on the nonlinear activation functions used to produce the outputs of the units in the neural network. Recently, the usage of bottleneck DNN has shown improved results in auto-encoder to reconstruct the input features [93]. In this research, the transformed features are investigated further and used to classify speaker age and gender.

In this section, the phoneme label extraction and the bottleneck (BN) extractor are introduced. Firstly, the labels are extracted for each frame for all utterances. Then based on the extracted labels, the BN extractor generates the T-MFCCs using a bottleneck layer in a trained DNN.

3.1.1.1 Pre-processing

Voice activity detection (VAD) is an essential step in most speech signal processing applications especially if background noise is present. The importance of VAD is due to the fact that it improves the speech intelligibility and recognition. Since the speech utterances used in this work were recorded in a public telephone center, the recorded utterances were exposed to noise and other interferences. As a result, VAD algorithm is necessary to reduce background noise and silent epochs in utterances to prepare them for feature extraction. In addition, cepstral mean variance normalization (CMVN) is applied to remove convolutional distortion and the linear channel effects. CMVN can be applied globally or locally. In this work, it is applied globally to get a normal distribution with zero mean and unit variance.

MFCCs set is one of the most well-known spectral feature sets and has been widely used in many speech applications. In this work, MFCCs set is employed. Figure 2. represents the flow diagram of MFCCs calculation.

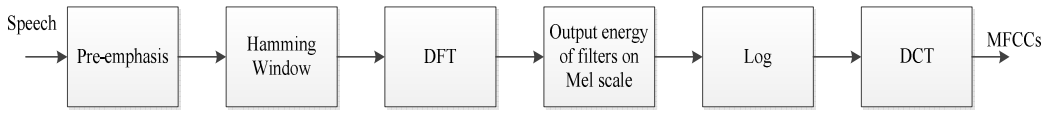


Figure 2. Mel-frequency cepstral coefficients.

The window size was chosen as 25 ms which is in the range of 20-40 ms per frame. This window duration was chosen to ensure the quasi-stationarity of the speech signal. Window size has a considerable effect on cepstral coefficients. If the window size is less than two pitch periods long, the cepstral coefficients may not show periodicity in the spectrum. At least two clearly defined periods should remain in the windowed speech segment [94]. In nature, the properties of speech signals change rapidly over time. Discrete Fourier transform (DFT) is used to calculate the power spectrum of each frame. A narrow mel-frequency filter bank is used for low frequencies while a wide mel-frequency filter bank is used for high frequencies. The main point of using the mel-frequency filter bank is to determine the energy level of different frequency ranges. In order to model the human ear, the log is taken for the filter bank energies. The discrete cosine transform (DCT) of the outputs of the log filter bank are calculated. In this work, speech utterances were divided into frames with 25 ms window size. 12 MFCCs and a normalized energy with their first and second derivatives (Δ's and ΔΔ's) were calculated for each frame, resulting in 39 coefficients representing each frame.

3.1.1.2 Phoneme label extraction (Tied-State Triphones)

Usually each database has a transcript file for each utterance that contains spoken words. Using the transcript along with speech audio files, the phonemes are extracted and this process is called grapheme-to-phoneme phase. The primary function of the HTK toolkit is to build Hidden Markov Models (HMMs) for speech-based tasks such as recognizers [95]. In the field of speech recognition, the recognition of speech is performed by mapping the sequence of speech vectors to the desired symbols sequence. Several complications may occur while performing the recognition of speech. For example, the

mapping between symbols and speech is not one-to-one. In most cases, the speech vector could be mapped to many symbols. Another complication is unclear boundary locations between words in a speech. This will cause incorrect mapping between the speech and the symbols. HTK tool is designed to address such issues using HMMs. HMMs are used to align phonemes with correct labels. It provides word isolation to deal with the unclear boundary location problem. In this work, the HTK tool in [95] is utilized to find the tied-state triphones which will be used later as labels for the output layer in the DNN.

The steps of finding the tied-state triphones is depicted in Figure 3 and described below.

Step 1: Generate the monophones by considering all of the pronunciations of each utterance in the database. The pronunciation that matches the best to the speech audio will be selected as an output.

Step 2: Produce triphones. Monophones are used to produce triphones. The current monophone, X, the previous monophone, L, and the next monophone, R, are processed together.

Step 3: Generate triphones that do not exist in the training data. These are called tied-state triphones.

Step 4: Find the best match between each frame of the speech utterance and tied-state triphones. The best match will be the phoneme label of the corresponding target frame.

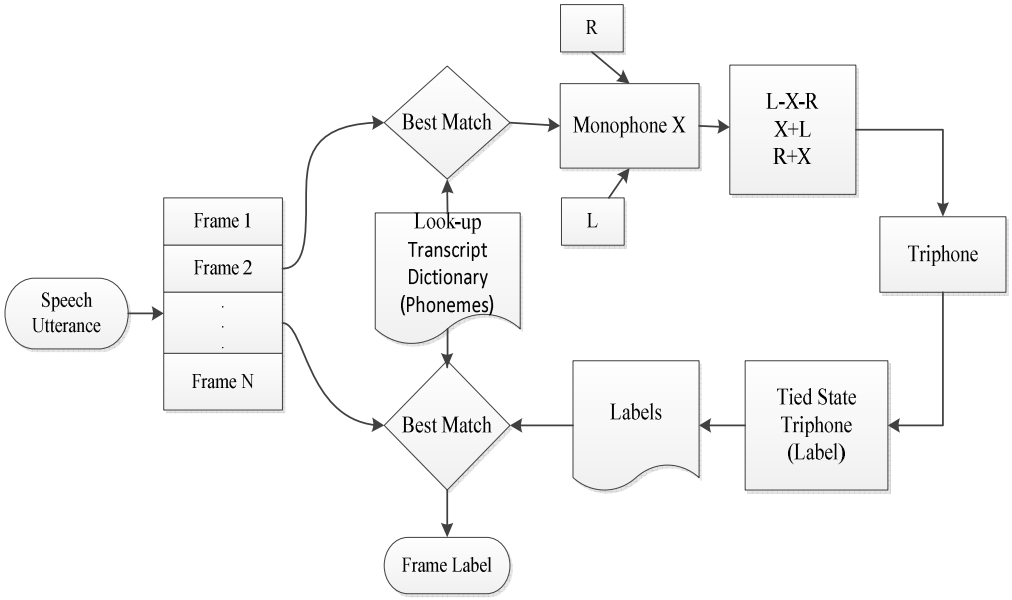


Figure 3. HTK process for extracting phoneme frame labels.

The phoneme labels are used for speech recognition. In this work, the phoneme labels are used to create transformed features. It keeps the phoneme specific characteristics of each speaker. The phoneme labels also help the DNN to embrace distinctive information in the transformed features.

3.1.1.3 Transformed features extraction

In this section, the process of extracting the transformed features is discussed. First the DNN training procedure is performed in two phases: the generative (unsupervised) and the supervised. Then, the process of extracting the transformed features based on the trained DNN will be explained in the BN extractor section.

A) DNN training

The first phase is generative. The DNN is pre-trained by using an unsupervised learning technique that employs the RBM. The second phase is discriminative. The DNN is trained by using the back-propagation algorithm in a supervised way. An RBM has input layer, V (visible layer) where V = {v_1, v_2, ..., v_V}, and the output layer, H (hidden layer) where h = {h_1, h_2, ..., h_H} [96]. The visible and the hidden layers consist of units. Each unit in the visible layer is connected to all units in the hidden layer. The restriction of this architecture is that there is no connection between the units in the same layer. Two types of RBMs, BB-RBM and GB-RBM [97] are used in this work. In the BB-RBM, the visible and hidden layer unit values are binary, $V \in \{0,1\}$ and $H \in \{0,1\}$. The energy function of the BB-RBM is defined in Equation (1)

$$E(v,h) = -\sum_{i=1}^{V}\sum_{j=1}^{H} v_i h_j w_{ij} - \sum_{i=1}^{V} v_i\, b_i^v - \sum_{j=1}^{H} h_j b_j^h \quad (1)$$

where V_i is the visible unit in layer i and H_j is the hidden unit in layer j. W_{ij} denotes the weight between the visible unit and the hidden unit. b_i^v and b_j^h are the bias of the visible unit in layer i and the hidden unit in layer j, respectively. For the GB-RBM, the visible unit values are real, where $V \epsilon R$, and the hidden units values are binary, where $H \epsilon \{0,1\}$. The energy function of this model is defined as in Equation (2)

$$E(v,h) = -\sum_{i=1}^{V}\sum_{j=1}^{H} \frac{v_i}{\sigma_i} h_j w_{ji} + \sum_{i=1}^{V} \frac{(v_i - b_i^v)^2}{2\sigma_i^2} - \sum_{j=1}^{H} h_j\, b_j^h \quad (2)$$

where σi is the standard deviation of the Gaussian noise for the visible unit i. The joint probability distribution which is associated with configuration of (v, h) is defined in Equation (3)

$$p(v, h; \theta) = \frac{\exp(-E(v, h; \theta))}{Z} \quad (3)$$

θ represent the weights and the biases, while Z is the partition function defined as in Equation (4).

$$Z = \sum_{v} \sum_{h} \exp\left(-E(v, h; \theta)\right) \quad (4)$$

The RBM is the basic building block in DBN. It is used as a feature detector and trained in an unsupervised way. The output of a trained RBM is used as an input to train another RBM. Training RBM is very useful for complex problems where the structure of the data is complicated and the implicit features could not be detected directly [98]. A number of RBMs could be stacked together to represent complex structures and to detect implicit features from the previous RBM representation in the stack. The stacked RBMs represent a generative model called DBN. The learning algorithm in the DBN is layer-wise and unsupervised. The layer-wise learning helps to find descriptive features that represent correlation between the input data in each layer [99]. The DBN learning algorithm works to optimize the weights between layers. Moreover, it is proved that initializing the weights between layers in the DBN network enhances the results more than if random weights are used. Another advantage of DBN training lies in its ability to reduce the effect of over-fitting and under-fitting problems where both are common problems in models with big number of parameters and deep architectures. After the DBN learning is completed and the weights between the layers in the DBN stack are optimized, the supervised training process is started by adding a final layer of labels on top of the DBN layers. These labels represent

the final classes of the whole network. In our work, these labels represent the tied-state triphones for the utterance speech data.

B) BN extractor

BN extractor architecture is generated from a trained DNN where each layer represents a different internal structure of the input features. In the DNN, the output of each hidden layer produces transformed features. All the layers above the bottleneck layer are removed to produce the BN extractor as shown in Figure 4. Figure 4 explains the proposed bottleneck DNN architecture using the phoneme labels. Figure 4 (a) explains the pre-training phase in the DBN consisting of five RBM layers. The first layer is a GB-RBM and the rest are BB-RBM with the bottleneck layer located in the middle. Figure 4 (b) portrays the DNN architecture which is formed by adding a softmax output layer on top of the DBN architecture. The weights for the DNN are tuned during supervised phase.

Introducing bottleneck layer has many benefits as reducing the number of units inside the bottleneck layer, getting rid of redundant values from the input feature set, and reflecting the class labels during the classification process [100-101]. It also helps to capture the descriptive and expressive features of short-time speech utterances [102]. Given a BN extractor with M layers, the features at the output layer can be extracted using Equation (5).

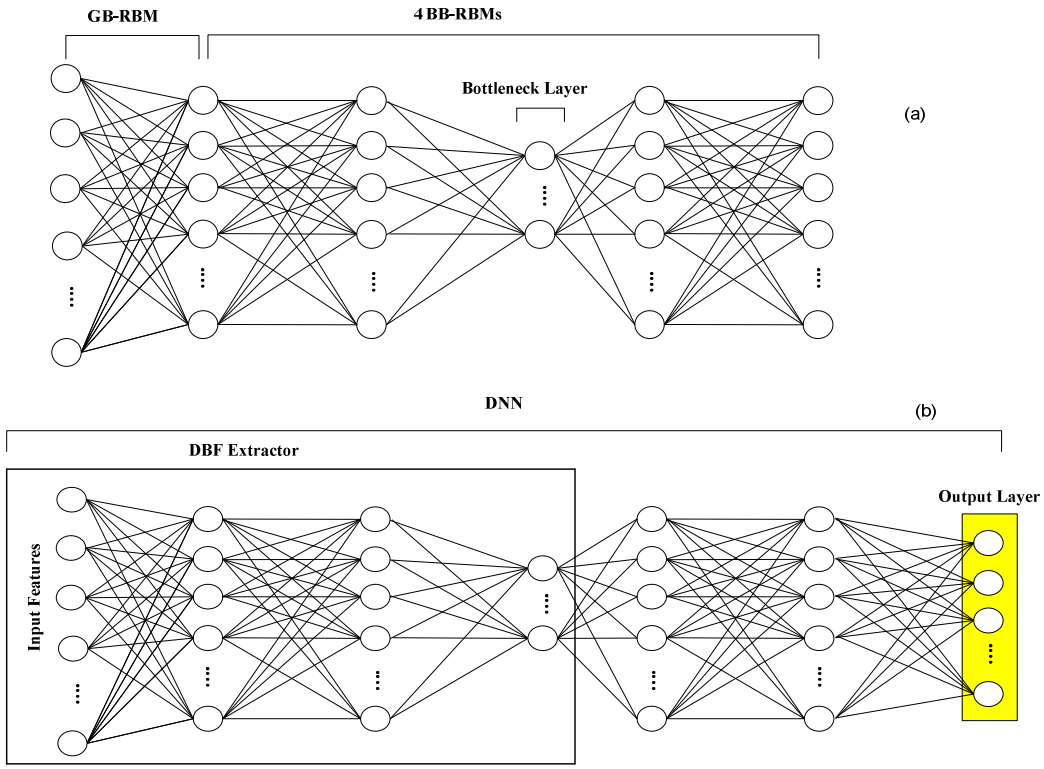


Figure 4. The process of DBF extracting using DNN (a) Pre-training (unsupervised phase). (b) Fine-tuning (supervised phase).

$$\begin{cases} l_1(x) = \sigma\left(\sum_{n=1}^{N} \mathrm{w}(x_n + b_1)\right) \\ l_2(x) = \sigma\left(\sum_{n=1}^{\mathrm{F_2}} \mathrm{w}(x_n + b_2)\right) \\ \vdots \\ l_M(x) = \sigma\left(\sum_{i=1}^{\mathrm{F_M}} \mathrm{w}(l_{m-1}(x) + b_M)\right) \end{cases} \tag{5}$$

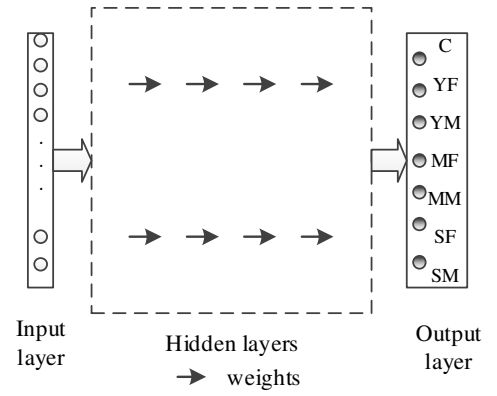
where σ is computed by the logistic function $\sigma(\mathrm{x}) = 1/(1 + \exp(-\mathrm{x}))$. $X = \{X_1, \ldots, X_N\}$ is the feature set vector, and N is the number of input features. L_M is the output of the M^{th} layer. F is a varying number that represents the input for each layer in the BN extractor. w

represents the weights between the input and output nodes in each layer. *B* represents the bias for each layer.

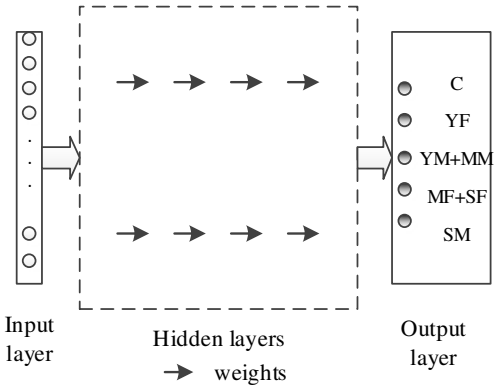
3.1.2 Regularizing DNN weights using shared class labels

Traditionally, one label is assigned to each class during the regularization of weights. However, in this work one label is allowed to represent two classes. Those two classes sharing the same label are chosen among the most misclassified classes. By sharing the same label, the weights between the DNN layers are being enforced to converge to an unbiased form with a wider-range representation. Misclassifications between classes are determined by a DNN classifier (Figure 5A). Two classes having the highest misclassification ratio are chosen to share a label. Let us have a database with seven classes, and the highest misclassifications occurred between classes (3 and 5), and between classes (4 and 6).

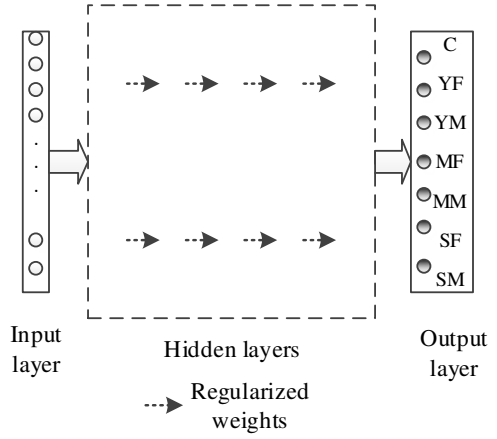
Therefore, five shared labels are generated, the first label is for the class 1, the second label is for the class 2, the third label is a shared label between the classes 3 and 5, the fourth label is shared between the classes 4 and 6, and finally the fifth label is for the class 7. As shown in Figure 5B a second DNN structure calculates the regularized weights. These regularized weights are used as initial weights for the third DNN classifier as shown in Figure 5C.



(A)



(B)



(C)

Figure 5. Shared Labels Method. (A) Finding misclassified classes. (B) Training a second DNN with shared class labels to calculate regularized weights. (C) Initializing a third DNN with regularized weights.

3.2 Classifying Age and Gender Based on DNN Speaker Models

Typically, representing each class in age and gender classification relies on finding a general model that can capture the common characteristics of all speakers' age and gender information. In this paper, a model for each speaker in a class is built. The purpose behind this idea is to find the specific identity and concentrated characteristics of each speaker separately in order to minimize any loss of unique information related to any speaker. Since the core of this work relies on creating a model for each speaker, it is reasonable to work on a feature set that is proved to be successful in the field of speaker recognition. Motivated by the success of SDC in many speech processing fields, especially in speaker recognition, this work uses the SDC as the main feature set.

Age and gender classification problem consists of M classes, where each class has N number of speakers sharing the same age range and gender. The DNN is trained with NxM labels. The settings for the training process are given in the experimental section. After the DNN is trained, NxM speaker models are developed as shown in Figure 6.

Each model accumulates the output layer posteriors. The accumulation of each model is done by performing feedforward on the input set until the posteriors are computed for each speaker. Then, the accumulated posteriors of the output layer are normalized (L2 normalization) and averaged for each speaker as shown in Figure 7. As a result, each class will have N speaker models.

During the testing, a model will be created for the corresponding utterance using the same steps applied to build speaker models as shown in Figure 7. The cosine distance is calculated between the test utterance model and every speaker model. The similarity

between the test utterance and each class is computed by averaging the results of cosine similarity (*Sim*) between the test utterance and the speaker models belonging to the same class. Finally, the maximum similarity between the test utterance and each class is taken as the finale similarity score *S* as in Equation (6).

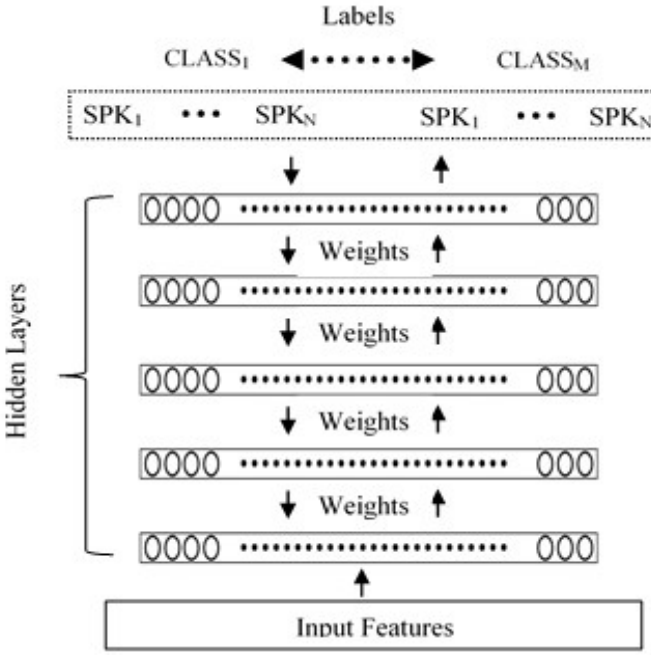


Figure 6. Speaker models, labels are N*M where M is the number of classes and N the number of speakers/class.

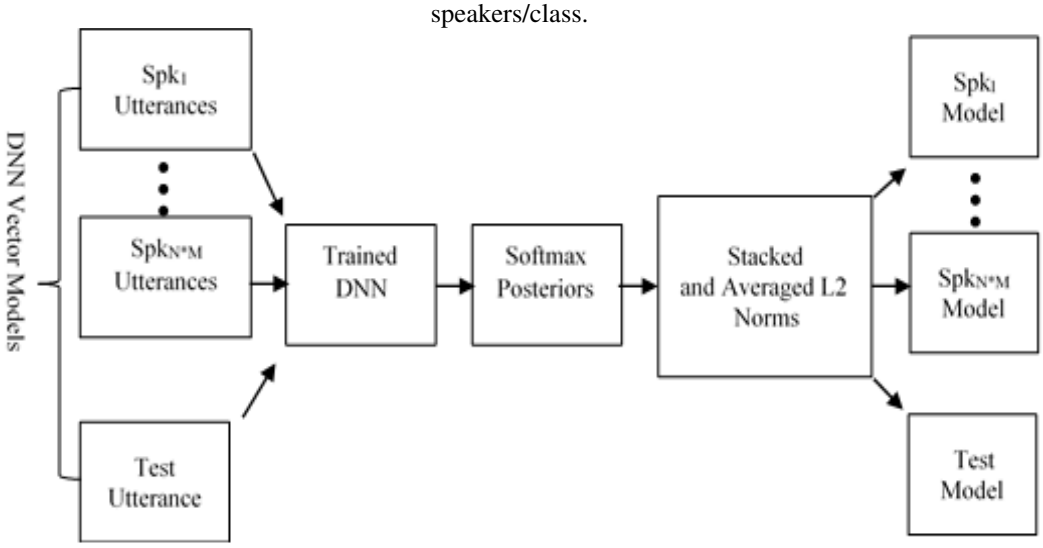


Figure 7. Flowchart of the proposed work, each class has N speakers.

$$S = \mathrm{Max}\left\{ \begin{matrix} \mathrm{M} \\ \mathrm{Sim_{cj}} = \mathrm{Avg}\left(\begin{matrix} \mathrm{N} \\ \mathrm{Sim(c_j Spk_i, Test_{utt})} \\ \mathrm{i} \end{matrix} \right) \\ \mathrm{j} \end{matrix} \right. \quad (6)$$

where *Avg* is the statistical average function, Sim_{cj} is the cosine similarity between the test utterance and the class *j*, $c_j Spk_i$ is the speaker *i* model of class *j*, and $\mathrm{Test_{utt}}$ is the test utterance model.

The two output vectors for a given test utterance are represented as *p* and *q*, where *p* and *q*, the output posterior probability of SDC Speaker Models (SSM) and SDC Class Models (SCM), are fused based on Equation (7).

$$S_j = \alpha p + (1 - \alpha) q \quad (7)$$

The final scoring for the corresponding utterance represents the index of the maximum value of the vector S_j. α is a parameter used to control the output result of the two models, and its value is set based on the performance of each model.

3.3 Jointly Fine-Tuning Age-Based DNN and Gender-Based DNN for Age and Gender Classification

The supervised training in DNNs aims to learn the optimal weights that will make the DNN classification process accurate with minimal overfitting. In this work, the supervised learning is divided into three parts; a DNN that learns the speakers age, a DNN that learns the speakers gender, and AGender-Tune that learns speakers age and gender together.

3.3.1 Gender-based DNN

This network is dedicated to capturing the gender of each speaker. As shown in Figure 8 the input for this network is the MFCCs set, and the output labels are Male and Female. The number of hidden layers is 5, where the number of nodes in each layer is 1024 node. Extracting speaker gender is easier than extracting the age or age and gender of the speaker. The achieved accuracy of Gender DNN is expected to reach high scores, and this will make the Gender DNN participation in other DNN networks effective.

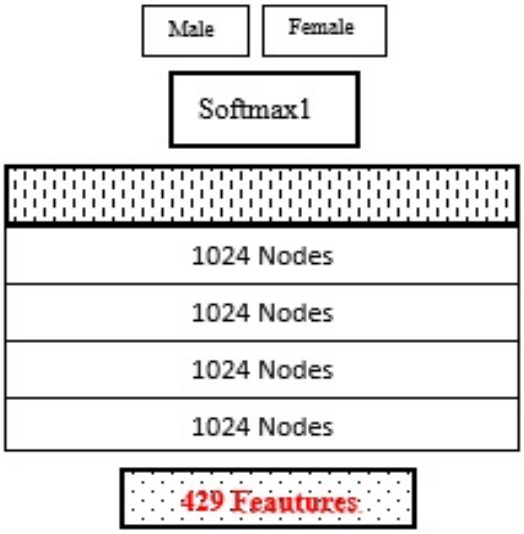


Figure 8. Gender DNN architecture.

3.3.2 Age-based DNN

This network will learn the speaker's age, where the input is the MFCCs feature set, and the output labels are children, young, mature, and senior. As shown in Figure 9, the number of hidden layers is five each consists of 1024 nodes. Decreasing the number of labels helps the classifier to achieve better results, the gender labels are separated from age labels to enable the classifier to focus on age prediction. In speech processing, it is known

that age classification is harder than gender classification, Age DNN will be trained to focus and learn as much as possible about speakers age, then Age DNN will be involved in a third DNN that utilize it.

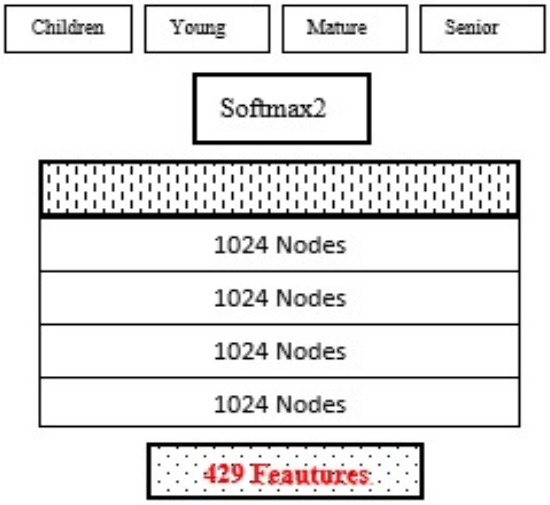


Figure 9. Age DNN architecture.

3.3.3 AGender-tune system based on DNN

In classification problems, two or more methods can be combined and utilized by fusing their results on the score level, but in these cases, the fusion may not utilize the full ability of each network. In this paper, an alternative way to combine two or more networks by fine-tuning their last hidden layers' outputs is proposed. Before combining, each network will be trained separately to utilize the network maximum ability.

First, to generate the new proposed AGender-Tune network, the two-trained age and gender networks are reused as shown in Figure 10. Next, a new output layer with a softmax activation function (softmax3) is added above the last hidden layers of both networks to jointly fine tune them together. The input for the newly added output layer is

the element-wise summation of the last hidden layer outputs of age network (O_1) and gender network (O_2) as in Equation (8).

$$X = O_1 \oplus O_2 \tag{8}$$

where $\oplus$ is the element-wise summation between each element in the two output vectors.

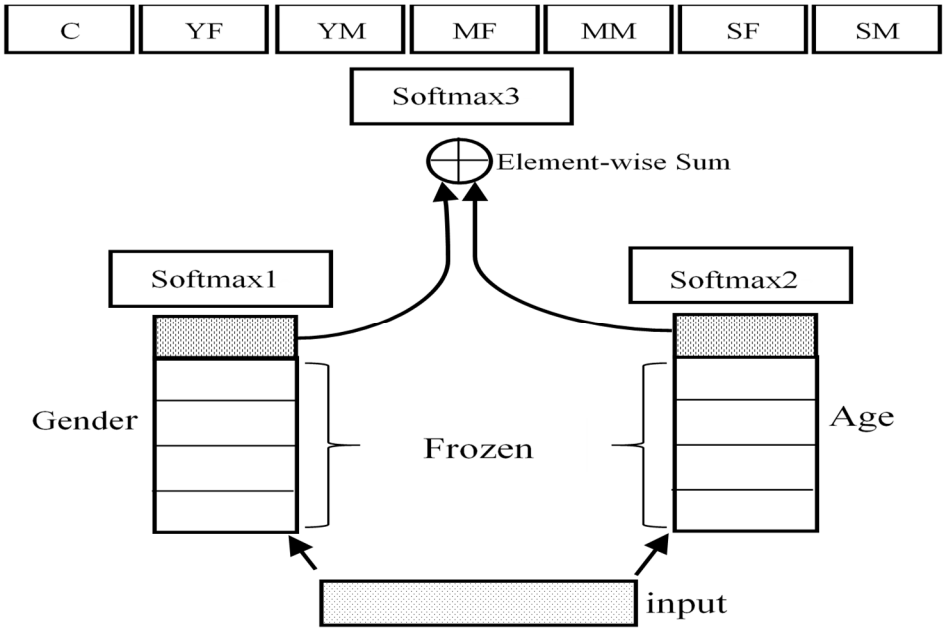


Figure 10. AGender-Tune network.

The output labels are 7, where each label represents a class for a group of speakers who share the same range of the age and gender. To combine the two networks, the weight values of the pre-trained age and gender networks are not changed (frozen). Then, the weight values of the last hidden layers of the Age, Gender, and the newly added output layer are trained and tuned as follows:

1) The newly added output layer is trained using softmax3. Consequently, the back-propagation process will take effect on the last hidden layers for the Age DNN and Gender DNN.

2) Whenever Age DNN receives updates from the newly added output layer, it starts updating its last hidden layer weights one more time using softmax1.

3) The same will be done for Gender DNN, whenever Gender DNN receives updates from the newly added output layer it; starts updating its last hidden layer weights one more time using softmax2.

4) Steps 1 to 3 are repeated until there is no learning gain.

5) Finally, after training is done, the final result (*S*) of speaker's age and gender classification are considered by taking the max of the newly added output layer (softmax3) as in Equation (9).

$$S = \text{argmax } O_{softmax3} \tag{9}$$

where $O_{softmax3}$ is the output posteriors of the newly added output layer (softmax3).

CHAPTER 4: EXPERIMENTAL SETTINGS AND DNNs CONFIGURATIONS

Several experiments have been conducted to evaluate the proposed methods and techniques. A publicly available database of speech utterance is used to compare the performance of the proposed work and compare it with other related work. The implementation code and simulation is written using MATLAB. The training of our networks is implemented using MatConvNet and DeepLearmMaster Toolboxs with some modifications. The computation time for training DNN depends on different factors: the size of the database (for speech utterances, there were millions of concatenated frames), the number of features for each sample, number of layers, and number of epochs. Therefore, the training has been conducted on a standard desktop with INVIDIA TITAN X with 12 GB. For all of our experiments, the utterance is divided into frames of 25 ms. In total, 39 features, one energy and 12- MFCCs or 12-SDC features with its first and second derivatives, are extracted for each frame. The number of nodes in the input layer is equal to the length of the input vector which has 39×n features. n is set to 11 after rigorous trial and error process. The 11 sequence frames are target frame and the previous and next (n-1)/2 frames. The training data is divided into mini batches. Each mini batch consists of 1024 utterances.

4.1 Database Specifications

Age-annotated database of german telephone speech (aGender) corpus, is used to test the performance of the proposed T-MFCCs using the GMM-UBM classifier. Each

speaker recorded six sessions using a mobile phone, the sessions were recorded indoor and outdoor to gain diverse environments. The utterances were sampled at 8 KHz and stored in 8-bit with A-Law format. The database consists of 47 hours of prompted and free text, which are command words, embedded commands, month, week day, relative time description, public holiday, birth date, time, date, telephone number, postal code, first name, last name, yes/no with according free or preset inventory and according 'eliciting' questions as "Please tell us any date, for example the birthday of a family member [103-104]. The number of speakers in the database is 954 and it includes seven categories of age and gender as shown in Table 1. The number of utterances in the database is 65364 and the average utterance length is 2.58 seconds, thus the utterances are considered as short utterances. The database was divided into two parts; the training set contains 53076 utterances (770 speakers) while the test set contains 17332 utterances (25 speakers/class).

Table 1. Age-annotated database of German telephone speech database.

Class	Category	Age Range	Gender	Abb.
1	Children	7-14	Male+Female	C
2	Youth	15-24	Female	YF
3	Youth	15-24	Male	YM
4	Adult	25-54	Female	AF
5	Adult	25-54	Male	AM
6	Senior	55-80	Female	SF
7	Senior	55-80	Male	SM

4.2 DNN Training Settings for Extracting The T-MFCCs Feature Set

Five hidden layers were used with 1024 nodes in each layer except the bottleneck layer where the number of nodes is 39. The number of nodes in the bottleneck layer is set to the number of input features, which is 39. The number of nodes in the output layer is set to the number of tied-state triphones, which is 4400, in the database. 10 epochs are used for training the GB-RBM over all the training data while 12 epochs are used for the rest of

the BB-RBMs. The learning rate for GB-RBM and BB-RBM is 0.0025. In the fine-tuning phase, five hidden layers of 1024 nodes each are used and 12 epochs are used. The learning rate is initially set to 0.1 for the first 6 epochs, and then it is decreased to one-half its initial value for the remainder epochs. The input data is the transformed features, while the number of epochs is 16. The learning rate was initially set to 0.1 for the first 3 epochs, then it is decreased to 0.8 times the old learning rate every two epochs. The momentum value was started at 0.5 for the first 3 epochs and then is increased to 0.9 for the remainder epochs.

4.3 DNN-Based Speakers Models Settings and Configurations

DNN architecture is used as a feature extractor and a classifier. Five hidden layers are used, and the number of nodes are 1024 in each layer. The number of output labels equals the total number of speakers in each class. In the training process, 12 epochs are used. The learning rate is initially set to 0.1 for the first 6 epochs, and then it is decreased to one-half its initial value for the remaining epochs.

4.4 DNN AGender-Tune System Training Settings

Age DNN and Gender DNN have the same network settings as the previous networks but they differ in the number of output labels. For Gender DNN, the output labels are Male and Female, while for the Age DNN the number of output labels are children, young, mature, and senior. When the two networks were fine-tuned into AGender-Tune network, 20 epochs were used for training and the initial learning rate was 0.1 for the first six epochs, and then decreased to one-half its original value for the remaining epochs.

CHAPTER 5: EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Classification Results Using the Proposed T-MFCCs Feature Set

Table 2 shows the overall classification accuracy by using the T-MFCCs is 56.13% and 58.89 % by the I-vector and DNN classifiers, respectively. On the other hand, the classification accuracies by using the traditional MFCCs are calculated as 43.60% and 45.89% by the same classifiers. The classification accuracies of MF, MM, SF, and SM classes are increased drastically. The T-MFFCs that are generated for the first time in this work increased the overall classification accuracy by about 13%. One of the reasons for this improvement is that the T-MFCCs features represent the prosodic features in addition to spectral features. The involvement of the phoneme labels in the generation of the T-MFCCs made it possible to grasp the prosodic features, such as intonation, stress, tone, and rhythm, of a speaker. Another reason is that the transformed features are the result of using phoneme labels in the training data, and this helped to remove any noise or silent frames so that the transformed features are calculated without acoustic background noise.

Figure 11. shows the receiver operating characteristics (ROC) of the transformed and traditional MFCCs (with random and regularized weights) by using DNN and I-vector classifiers. The ROC curves are calculated by using one-agianst-all rule. The area under curve (AUC) for the T-MFCCs is found to be bigger than the traditional MFCCs (Table 3 compares the AUC for both sets). The AUC values are calculated as in [105]. The DNN classifier performes better than the I-vector classifier in terms of AUC.

Table 2. The overall classification accuracies of the DNN and I-Vector classifiers using the traditional and the T-MFCCs (%).

Classifier		C	YF	YM	MF	MM	SF	SM	Ovll. Acc
I-vector	Traditional MFCCs	64.86	57.12	49.01	24.50	27.03	49.91	32.80	**43.60**
	T-MFCCs	60.33	66.49	48.00	45.46	48.56	56.89	67.15	**56.13**
DNN with regularized weights	Traditional MFCCs	54.33	52.60	44.80	25.13	42.33	46.13	55.87	**45.89**
	T-MFCCs	62.23	61.54	53.38	47.69	52.00	64.23	70.77	**58.98**
DNN with random weights	Traditional MFCCs	56.53	47.27	49.07	27.53	35.33	36.13	53.80	**43.67**
	T-MFCCs	59.69	60.15	48.85	40.08	52.23	60.92	63.38	**55.04**

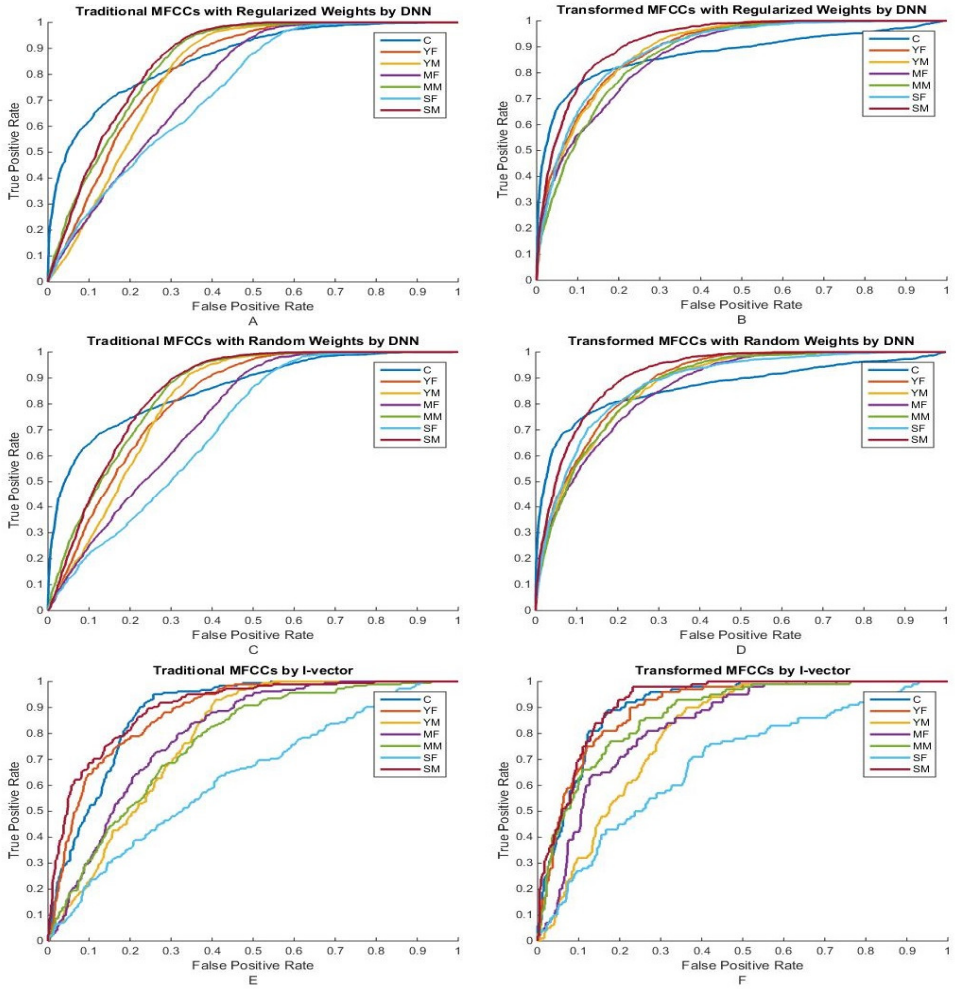


Figure 11. ROC curves of different classifier scenarios. A) The DNN with regularized weights and the traditional MFCCs. B) The DNN with regularized weights and the T-MFFCs. C) The DNN classifier with random weights and the traditional MFCCs. D) The DNN classifier with random weights and the T-MFFCs by. E) The I-vector classifier by using traditional MFCCs. F) The I-vector by using the T-MFCCs.

Table 3. Corresponding AUC measurements for classification of speaker's age and gender.

Class	DNN Regularized Weights		DNN Random Weights		I-vector	
	Traditional MFCCS	T-MFCCs	Traditional MFCCS	T-MFCCs	Traditional MFCCS	T-MFCCs
C	0.86	0.87	0.86	0.87	0.88	0.90
YF	0.81	0.89	0.82	0.88	0.88	0.89
YM	0.81	0.89	0.81	0.87	0.78	0.80
MF	0.76	0.87	0.75	0.86	0.79	0.83
MM	0.85	0.87	0.85	0.87	0.76	0.87
SF	0.74	0.89	0.71	0.88	0.63	0.68
SM	0.86	0.92	0.85	0.91	0.89	0.92
Overall AUC	**0.81**	**0.89**	**0.80**	**0.88**	**0.80**	**0.84**

Another analysis to compare the T-MFCCs and the original MFCCs was done by comparing variations between standard deviation of the 12 MFCCs plus a normalized energy parameter for each class for additional insight. It is shown in Figure 12. It is observed that T-MFCCs features present less intra-class variation than the original MFCCs. It is also observed that there is significant inter-class variation in the T-MFCCs features. Minimal intra-class variation and maximal inter-class variant in features are preferred in order to have better classification. The improvement in the classification of the adult male class is not significant. The speakers in this class are misclassified as young male or senior male. The statistical analysis of the original MFCCs and T-MFCCs shows that the distribution of the first thirteen cepstral coefficients among the male classes is similar in terms of standard deviation (Figure 12). Although T-MFCCs are presented more uniformly intra-classes compared to the original MFCCs for all male speakers, they also have similar distribution inter-male classes. That is why no significant improvement observed in age and gender classification of male speakers.

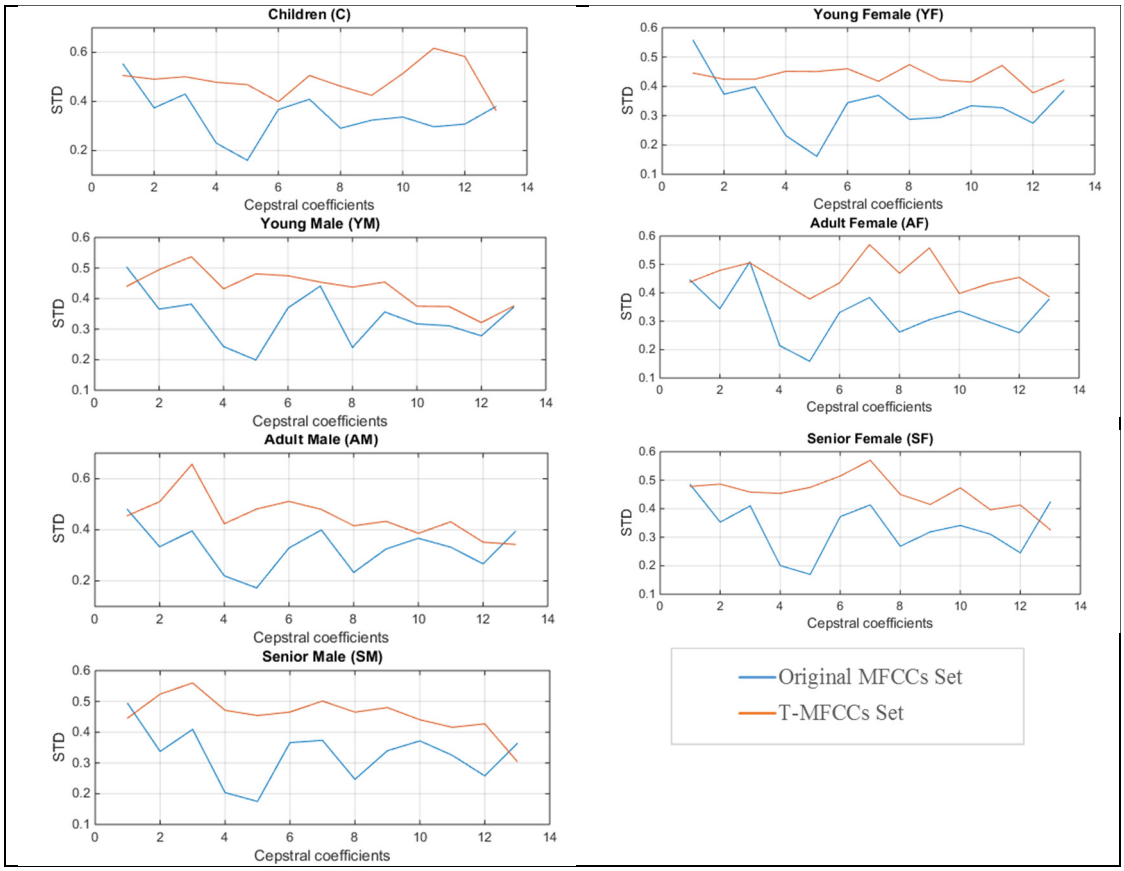


Figure 12. Variation between standard deviation values of the first 13 coefficients of the original and T-MFCCs sets for all classes.

For more clarification as shown in Figure 13 and 14 the T-MFCCs features have good variance among female classes. It is reflected as an increase in the classification accuracies for female speakers. On the other hand, it is observed that the MFCCs and T-MFCCs features have less variance among the male classes compared to that of female classes. As a result, misclassification occurred among male classes, especially between young- and adult male classes.

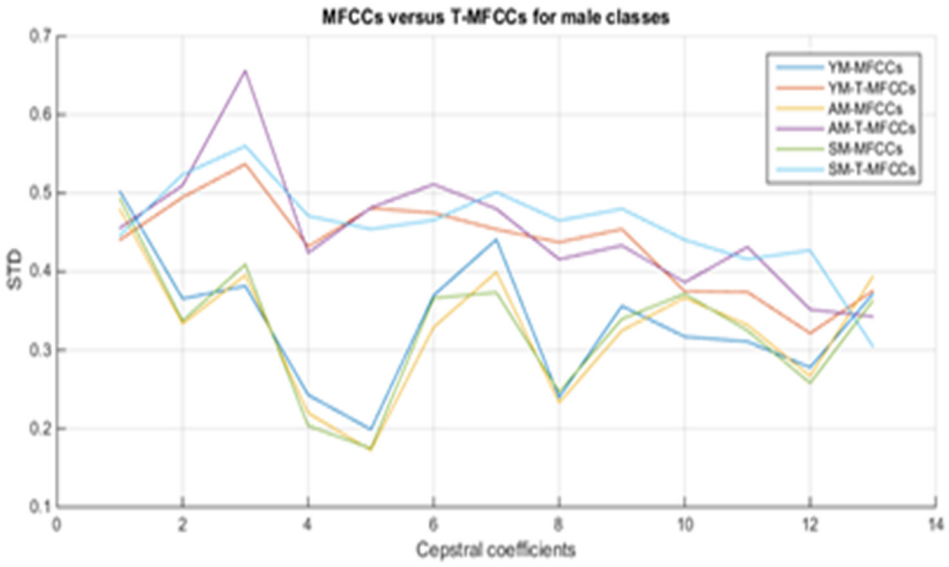


Figure 13. MFCCs versus T-MFCCs sets for all male classes.

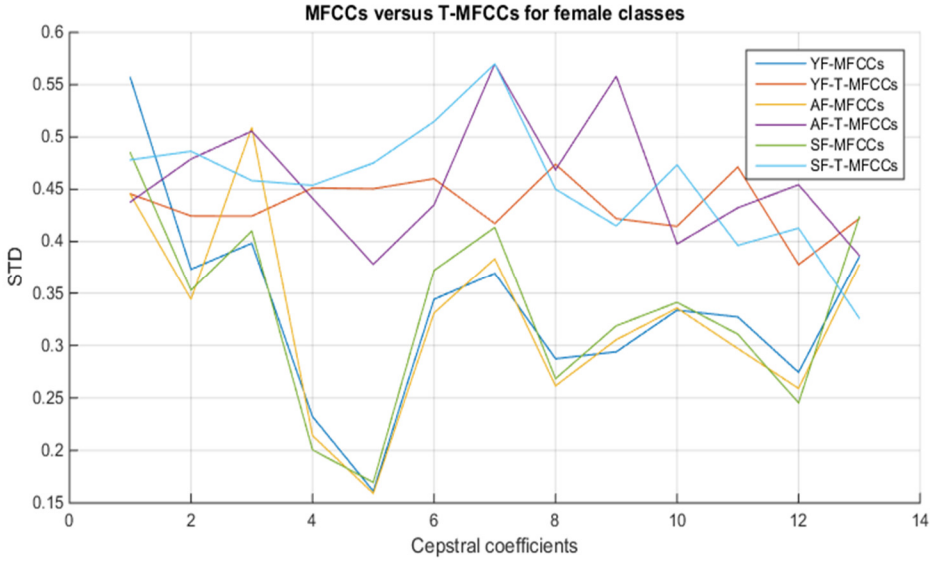


Figure 14. MFCCs versus T-MFCCs sets for all female classes.

As comparing the classifiers, the DNN classifier performed slightly better than the I-vector classifier. Figure 15, shows the variance in weights at each layer in the DNN classifier by using random weights and regularized weights. Higher variance between the weights in each layer is needed to distinguish different classes. As it can be seen in Figure 15, the variance between the weights using shared labels is higher than that of the randomly

initialized weights, therefore, the regularized weights converge faster than the random weights for most of the DNN layers.

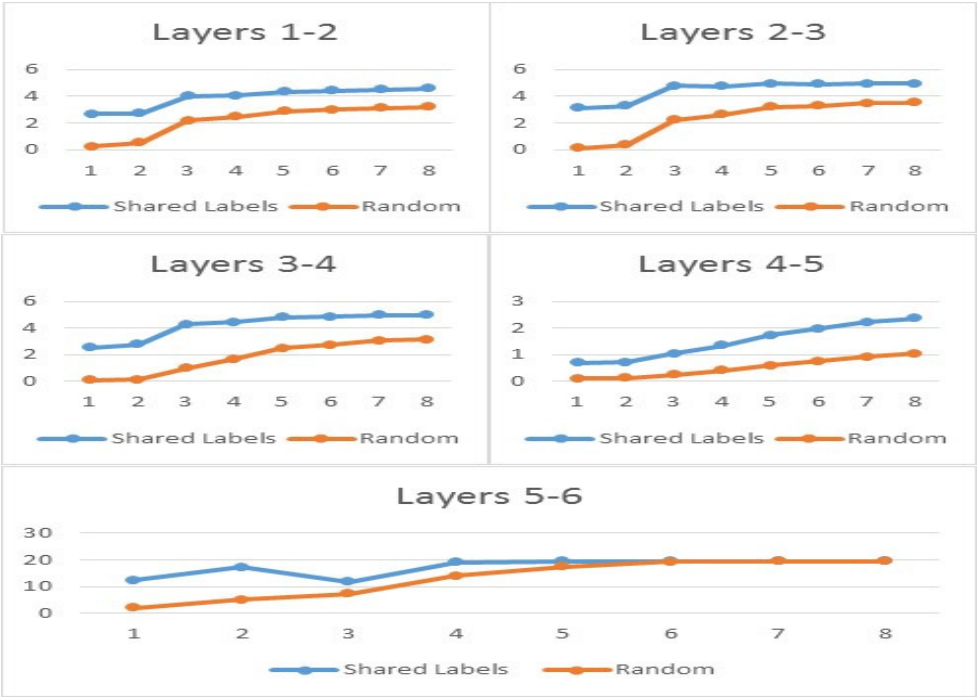


Figure 15. Variance versus epoch number graphs of regularized and random weights between layers. The x-axis represents the epoch number (1-8), and y-axis represents the variance (y is scaled by 1000).

Table 4 presents the confusion matrix by using the I-vector classifier. It can be seen that children (C), young female (YF), and senior (SM, SF) classes are classified with higher accuracies compared to the other classes. The major classifications occurred among the same-gender classes. Young female (YF) and senior male (SM) classes have the highest accuracy rates and are correctly classified as 66.49% and 67.15%, respectively. Middle and senior female groups (MF, SF) are classified with the accuracy of 45.46% and 56.89%. Children (C) and young male (YM) classes achieved the accuracy of 60.33% and 48%.

Table 4. Confusion matrix of the I-vector classifier using the transform MFCCs set (%).

Predicted / *Actual*	**C**	**YF**	**YM**	**MF**	**MM**	**SF**	**SM**
C	**60.33**	27.90	1.5	4.80	0	2.88	2.59
YF	21.08	**66.49**	2.70	6.85	0	2.88	0
YM	8.89	1.62	**48**	0.18	18.97	10.99	11.35
MF	3.60	16.85	2.52	**45.46**	2.16	29.23	0.18
MM	3.42	1.26	24.43	4.14	**48.56**	2.34	15.85
SF	7.41	11.17	5.23	13.18	1.44	**56.89**	4.68
SM	4.50	0.72	11.53	0	15.56	0.54	**67.15**

Table 5 and Table 6 present the confusion matrices of the DNN classifier using the transformed and traditional MFCCs with regularized weights. In Table 5, the class SM is classified with the highest accuracy (70.77%), while the classes YF, C, and SF are correctly classified with the accuracy ranges between 61% and 64%. The classification accuracies of the MM and YM classes are calculated as 52% and 53.3%, respectively. The lowest accuracy was achieved by the class of MF, as 47.69%. It is observed that the highest misclassification rates have always occurred between the classes with the same gender and close age, or between the children and young female class.

Table 5. Confusion matrix of the DNN classifier using the transform MFCCs set (%).

Predicted / *Actual*	**C**	**YF**	**YM**	**MF**	**MM**	**SF**	**SM**
C	**63.23**	15.38	4.08	3.31	5.08	4.54	4.38
YF	15.92	**61.54**	0	11.08	0.54	10.23	0.69
YM	1.62	0.62	**53.38**	2.38	24.46	2.15	15.38
MF	3.38	16.08	2.15	**47.69**	0.77	28.85	1.08
MM	0.69	0.92	21.77	0.85	**52**	2.23	21.54
SF	4.69	8.92	1.77	16.23	0.923	**64.23**	3.23
SM	0.46	0.31	11.85	0.38	13.69	2.54	**70.77**

Table 6. Confusion matrix of the DNN classifier using the traditional MFCCs set (%).

Predicted / *Actual*	**C**	**YF**	**YM**	**MF**	**MM**	**SF**	**SM**
C	**54.33**	22.88	2.67	6.13	0.73	11.13	2.13
YF	13.00	**52.60**	0.40	16.47	0.20	16.93	0.40
YM	0.87	1.00	**44.80**	2.13	26.20	4.60	20.40
MF	4.40	26.47	1.73	**26.13**	1.53	37.67	2.07
MM	1.07	0.80	30.93	1.40	**42.33**	2.60	20.87
SF	4.27	16.20	3.00	23.27	1.20	**46.13**	5.93
SM	1.07	0.47	10.98	0.67	26.87	4.07	**55.87**

By comparing the classification accuracies of each class in Table 5 and Table 6, the T-MFCCs help to improve the DNN performance about 10% higher for the classes C, YF, YM, and MM and between 15-20% higher for the classes MF, SF, and SM. This observation can also be seen in the AUC measurements in the Table 3. In their work, Barkana and Zhou [106] reported that traditional MFCCs of the middle-aged female (MF) speakers and senior female speaker have very similar characteristics leading to misclassifications between these two classes. The proposed T-MFCCs decreased the misclassifications between these two classes significantly since phoneme labels are used in generating the transformed features. The transformed features contain phoneme specific characteristics of each speaker in addition to the spectral characteristics.

5.2 DNN-Based Speakers and Classes Models Results and Discussion

The overall classification accuracies for the MFCCs-Speakers Models (MSM), MFCCs-Class Models (MCM), SSM, SCM, and fused SSM+SCM are given in Table 7. The proposed SSM model achieved the best results among the other models.

The confusion matrices for the SCM, SSM, and fused SSM+SCM models are shown in Tables 8, 9, and 10. The confusion tables show that the highest misclassification rates occur between the same gender classes. In Figure 16 and 17, the performance of the young (Y), middle-aged (M), and senior (S) female and male classes for all models are compared, separately. It can be seen that all models achieved somehow poor results for MF and MM classes without the score level fusion. The SSM achieved the best result for these classes as 38.5% and 36.3%.

As shown in Figure 16, for the female classes, the SSM achieved the best results except for the YF class (56%), where SCM achieved slightly better result (57.4%). This result supports the effectiveness of the SDC feature set over MFCCs. The SSM outperformed the other models in male classes (Figure 17). In particular, SDC speaker and class models generated better classification results in female and male classes than MFCCs speaker and class models. However, a significant improvement (57.21%) is achieved when (SSM+SCM) models were fused. As it can see, the fused system outperformed other models in all classes.

Table 7. Classification accuracies (%).

	MSM	**SSM**	**MCM**	**SCM**	**Fused (SSM+SCM)**
C	56.6	58.5	57.4	60.5	74.3
YF	55.4	56	45.7	57.4	70
YM	45.1	49.9	44.3	48.3	55.4
MF	32	38.5	35.4	30.7	39.3
MM	34.3	36.3	33.8	35	39.8
SF	43.7	45.8	35.7	44.2	55.3
SM	49.3	60	49.4	57.6	66.3
%	45.2	49.3	43.1	47.7	57.2

Table 8. Confusion matrix for SCM (%).

Predicted / *Actual*	**C**	**YF**	**YM**	**MF**	**MM**	**SF**	**SM**
C	**60.5**	17.1	8.5	3.2	3.4	6.5	0.8
YF	23.8	**57.4**	0.6	8.8	0.1	8.9	0.4
YM	3.3	1.8	**48.3**	2.4	21.0	3.2	20.0
MF	12.2	23.4	1.1	**30.8**	0.8	30.4	1.3
MM	1.8	0.3	27.9	1.0	**35.0**	3.5	30.5
SF	14.5	17.7	0.8	19.3	0.4	**34.3**	3.0
SM	1.0	0.3	15.6	0.9	22.1	2.4	**57.7**

Table 9. Confusion matrix for SSM (%).

Predicted / *Actual*	**C**	**YF**	**YM**	**MF**	**MM**	**SF**	**SM**
C	**58.5**	18.4	8.9	3.9	3.4	5.8	1.0
YF	22.0	**56.1**	0.4	11.3	0.2	9.8	0.3
YM	2.3	2.1	**49.9**	2.3	17.3	4.7	21.4
MF	9.3	21.4	1.0	**38.6**	0.8	27.5	1.5
MM	1.6	0.5	26.7	1.8	**36.3**	3.9	29.3
SF	11.0	17.4	1.1	20.5	0.3	**45.8**	3.8
SM	0.6	0.2	16.9	0.9	19.3	2.0	**60.1**

Table 10. Confusion matrix for fused SSM+SCM (%).

Predicted / *Actual*	C	YF	YM	MF	MM	SF	SM
C	**74.3**	12.9	4.3	2.6	1.3	3.3	1.4
YF	11.8	**70.0**	0.3	12.1	0.1	5.6	0.1
YM	1.2	0.7	**55.4**	1.7	19.1	3.4	18.6
MF	8.2	24.3	0.8	**39.3**	0.3	26.4	0.7
MM	0.5	0.0	22.3	0.4	**39.8**	0.4	36.6
SF	8.5	11.6	0.6	22.3	0.9	**55.3**	0.8
SM	1.0	0.1	9.8	0.3	19.9	2.5	**66.3**

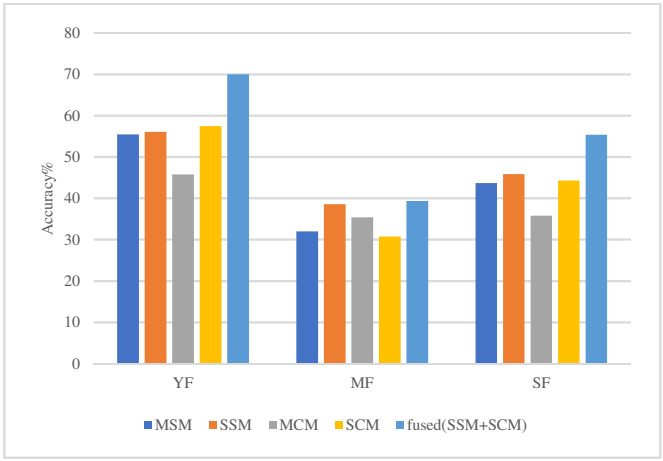


Figure 16. Comparison of classification accuracies between four methods for female speakers.

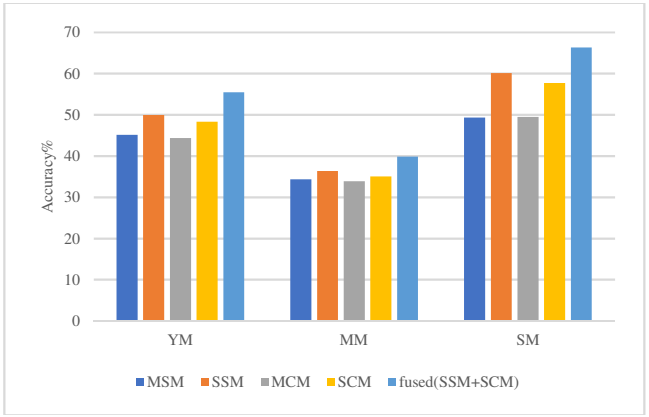


Figure 17. Comparison of classification accuracies between four methods for male speakers.

The performance of the fusion model α values is depicted in Figure 18. Several experiments are conducted to choose the optimal value of the α. The best performance occurred when α is 0.9.

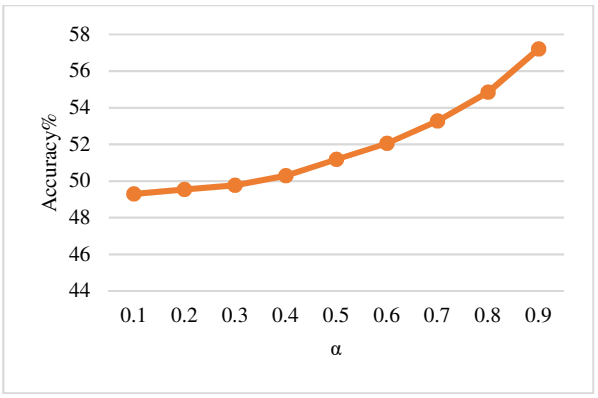


Figure 18. The performance of the fused (SSM+SCM) system with respect to the α values.

5.3 Results for DNN-Based AGender-Tune System

The classification accuracies are presented in Table 11. The proposed AGender-Tune system outperformed the GMM-UBM system by approximately 12% and outperformed the I-Vector system by almost 7%. The proposed work extracted the speaker's age separately from the gender before merging the last hidden posteriors of Age and Gender DNNs into one layer that is to be trained further. This separate pretraining helped to maintain the unique identity of each speaker even after age and gender posteriors became one layer.

The proposed system achieved a significant improvement especially in mature female and male classes (45.52%, 48.62%) and senior female and male classes (57.5%, 60.63%). As well as, the I-Vector classifier achieved better results than the GMM-UBM system in all classes except for the MM class. In children and SF classes, the proposed and the I-Vector systems achieved almost same results with a slight advantage for the AGender-Tune system. The confusion matrix for the proposed system is presented in Table 12.

It can be seen from Table 12 that the proposed system achieved a significant improvement for all classes especially for (MF, MM, SF, and SM). Our system was able

to discriminate between these classes better than the baseline systems. AGender-Tune system has been trained in two ways, first with separated age (Age DNN) and gender (Gender DNN) networks, second, with a shared output layer resulted from the Age DNN and Gender DNN output layers, and this shared output layer has seven age and gender labels.

Table 11. The classification accuracies of GMM-UBM, I-vector, and AGender-Tune system (%).

	GMM-UBM	**I-Vector**	**AGender-Tune**
C	55.6	64.9	65.7
YF	48	57.1	58.3
YM	41.9	49	49.9
MF	29.6	32.5	45.5
MM	41.2	36	48.6
SF	36.4	49.9	57.5
SM	53.9	45.8	60.6

Table 12. Confusion matrix for the AGender-Tune system.

Predicted / *Actual*	**C**	**YF**	**YM**	**MF**	**MM**	**SF**	**SM**
C	**65.7**	14.9	5.7	3.2	2.1	6.2	2.3
YF	11.8	**58.3**	0.5	19.9	0.7	8.3	0.5
YM	2.1	0.9	**49.9**	2.6	26.9	2.9	14.7
MF	8.6	18.2	1.2	**45.5**	1.2	24.8	0.5
MM	1.2	0.1	22.3	0.3	**48.6**	1.2	26.2
SF	8.1	9.1	1.2	21.1	1.5	**57.5**	1.5
SM	1.5	0	8.7	1.2	22.2	5.8	**60.6**

To evaluate the performance of the baseline systems and the proposed work when the time duration of the speech utterance is different; the overall accuracy of each system over five slots of time durations (1-5 seconds) were examined. Figure 19 shows the performance of the three systems. In general, the performance of all systems has been enhanced by increasing the duration of the utterance time; AGender-Tune system

performed better than the baseline systems for all time slots. A possible explanation refers to the fact that I-Vector and GMM-UBM systems could not build a good representation of eigenvector and GMM-UBM supervector for the corresponding utterance if it is short in time, and it is known that the aGender database utterances are short in time. When the duration of the utterance increases, for example from 3 to 4 seconds, the accuracy is not increasing for the GMM-UBM and AGender-Tune system, this is due to the sparse data of these utterances duration, where most of the YF, MF, and MM utterances exists in this time duration. Also, it is noticed from Table 11 that the higher misclassification occurs between these classes and these classes have the least accuracy results among other classes.

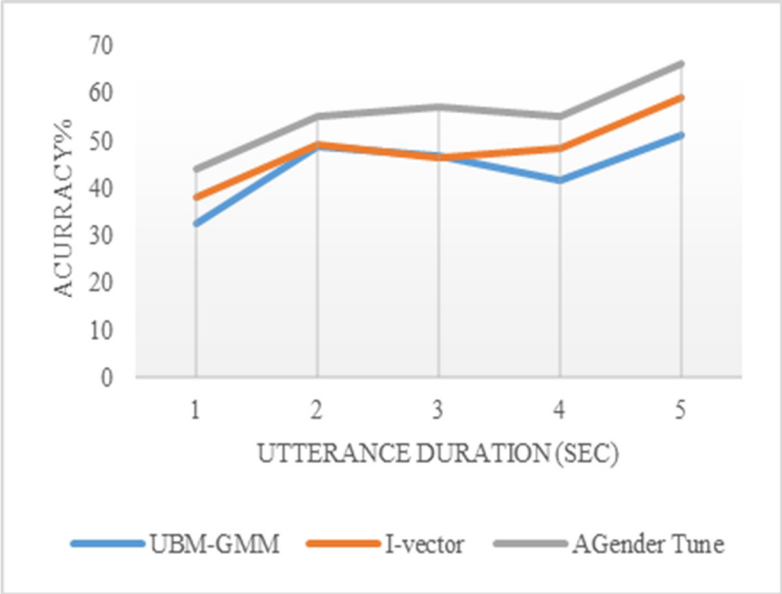


Figure 19. Comparison between the AGender-Tune and the baseline systems for different time duration utterances.

5.4 Fusion of the Speaker and Class Models Using the T-MFCCs Feature Set for Enhancing Speaker Age and Gender Classification

As shown in the previous sections, the proposed T-MFCCs feature set, the proposed speaker models, and the fusion of different combined systems have improved the classification accuracy for the speaker age and gender by a considerable margin. In section 5.2 several systems have been fused using the MFCCs feature set and the SDC feature set. The best results have been achieved by using the proposed speaker models, therefore, in this section the proposed T-MFCCs feature set will be used as input for a fused system which combines the speaker and class models.

As shown in Figure 20, the fused system consists of two DNN networks. The speaker models are used as labels on the left network, while the class models are used as labels on the right network. The input for both networks is the proposed T-MFCCs feature set. The training of both networks is discussed in sections 5.1 and 5.2. In the test phase, the fusion is calculated over the score level of both networks for the same utterance as discussed in section 3.2.1.

Table 13 shows the confusion matrix for the left network in Figure 20 which uses the T-MFCCs feature set as input and the speaker models as output labels. The results of the right-side network in Figure 20 are shown and discussed in section 5.1. The speaker models based network achieved 59.59% overall accuracy, while the class models based network achieved 58.98%. Comparing the results for both networks, both networks achieved good results for speaker age and gender classification with a slight advantage for the speaker models based network in terms of overall accuracy. As well as, comparing the

classification results for each class it can be observed that the speaker models based network performed better results for most classes than the class models based network.

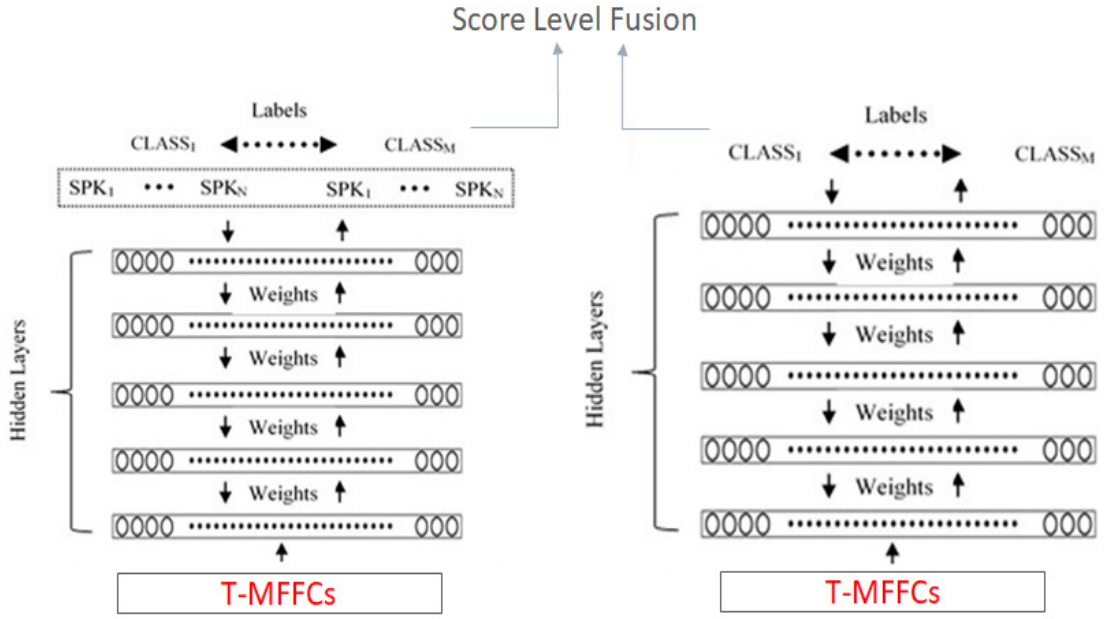


Figure 20. Score level fusion of speaker and class models using the proposed T-MFCCs.

Table 13. Confusion matrix for the speaker models using the T-MFCCs feature set.

Predicted / *Actual*	C	YF	YM	MF	MM	SF	SM
C	**65.73**	13.18	3.35	2.41	6.07	4.74	4.52
YF	12.85	**61.66**	1.52	12.14	1.01	9.93	0.89
YM	1.7	0.93	**50.47**	3.81	26.67	3.2	13.22
MF	2.28	18.03	1.6	**44.53**	1.02	31.12	1.42
MM	0.27	1.15	19.43	0.76	**54.36**	1.14	22.89
SF	2.91	7.36	0.86	18.97	0.42	**67.44**	2.04
SM	0.32	0.12	10.57	0.18	14.24	1.61	**72.96**
Overall Accuracy							**59.59**

The results for the proposed fused system is shown in Table 14. The overall accuracy for the fusion system is 61.16% which is better than the speakers and class models. The fused system achieved better results than the separated speakers and class models system for the C, YF, MM, and SM classes. In most cases, the fusion of two systems can improve the classification accuracy, and this can be observed in section 5.2. On the

other hand, utilizing the T-MFCCs as input feature set either for speaker or class models has gained significant advantage when compared with utilizing other feature sets. As a result, the fused speaker and class models system has a slight better result when compared with the result of each system.

Table 14. Confusion matrix for the score level fusion of the speaker and class models using the T-MFCCs.

Predicted / *Actual*	**C**	**YF**	**YM**	**MF**	**MM**	**SF**	**SM**
C	**69.52**	11.08	3.32	1.66	4.86	6.64	2.92
YF	14.09	**66.94**	2.01	8.19	0.85	7.45	0.47
YM	2.78	0.46	**48.36**	4.75	22.43	2.91	18.31
MF	2.45	20.57	1.3	**42.76**	1.85	29.14	1.93
MM	0.49	0.76	18.14	0.52	**59.39**	0.93	19.77
SF	1.99	11.48	1.06	17.59	0.63	**65.49**	1.76
SM	0.26	0.05	8.49	0.35	12.81	2.39	**75.65**
Overall Accuracy							**61.16**

5.5 Utilizing Speaker Models for the AGender-Tune System

In section 3.3, the proposed AGender-Tune System relied on the class models as output labels for classification. Since the speaker models have proved its efficiency for speaker age and gender classification as shown in section 3.2 and 5.2, In this section a new architecture is proposed by using the speaker models as the output labels. As shown in Figure 21, the left side represent the AGender-Tune system using the class models as output labels, and the right side shows the new architecture for the AGender-Tune system with speaker models as output labels. The system is trained and the test is carried out as discussed in section 3.3.

Table 15 shows the results of the AGender-Tune system using the speaker models as output labels. The table shows an improvement by approximately 2% in terms of overall accuracy compared with the AGender-Tune system using the class models as output labels.

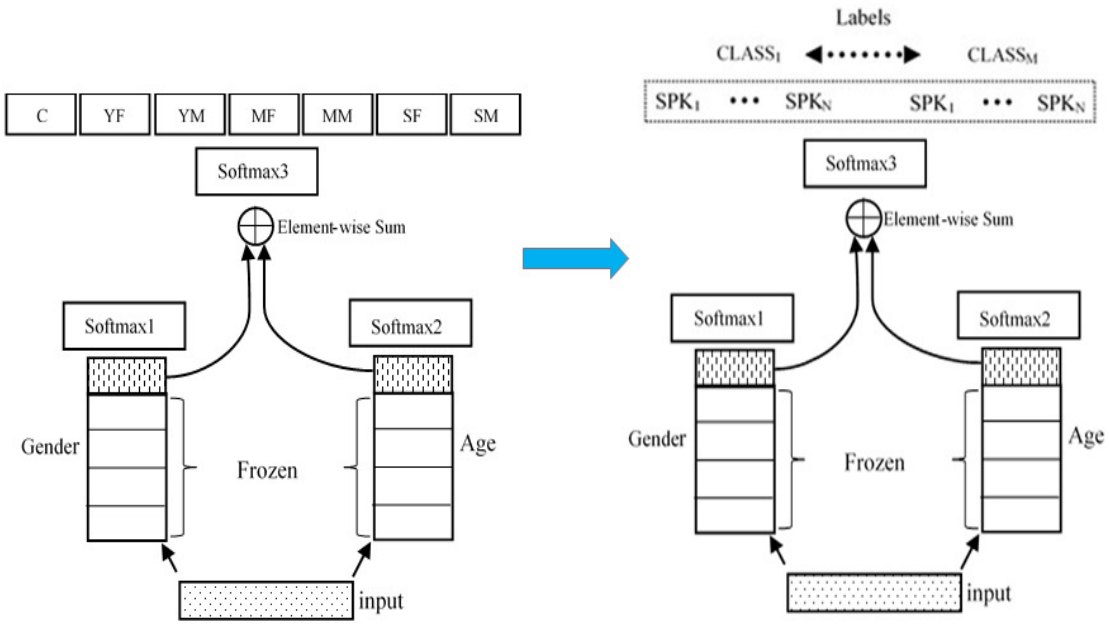


Figure 21. AGender-Tune system using the speaker models as output labels.

Table 15. Confusion matrix for the AGender-Tune system using the speaker models.

Predicted / *Actual*	**C**	**YF**	**YM**	**MF**	**MM**	**SF**	**SM**
C	**68.52**	18.24	2.89	2.16	1.37	4.96	1.86
YF	15.25	**61.12**	0.86	15.46	0.13	6.46	0.72
YM	1.71	0.24	**45.62**	1.82	30.07	1.47	19.07
MF	2.33	13.62	2.04	**52.23**	1.07	28.18	0.53
MM	0.72	0.38	15.61	0.65	**53.18**	0.13	29.33
SF	4.77	13.49	0.76	24.01	1.2	**54.94**	0.83
SM	0.42	0.11	5.33	1.05	23.16	2.66	**67.27**
Overall Accuracy							**57.55**

Comparing the fusion matrix of both systems, the AGender-Tune system based on speaker models as output labels achieved better results for most classes than the class models based system. These results support the efficiency of utilizing the speaker models based system for speaker age and gender classification.

5.6 A Comparison Between the Proposed Work and State-of-the-Art

The overall accuracies of the previous studies using the aGender database and MFFCs as input feature set are listed in Table 16. The classification accuracies for these systems are reported in [85, 88, 89]. The best result is achieved in [85] by fusing all their proposed systems together manually (MFuse 1+2+3+4+5).

From Table 16, it is noticed that the proposed methods achieved the highest classification accuracies when compared with state-of-the-art methods that works on the same database and classifies the same number of classes. The accuracy of the speaker age and gender classification is improved by approximately 9% when compared to the (MFuse 1+2+3+4+5) system. Our proposed fused model (6+9) system achieved the highest accuracy when compared with the rest of our proposed methods. In addition, the fused SDC-class and SDC-speaker model outperformed the best of the state-of-the-art (MFuse 1+2+3+4+5) system, even though each system alone has less classification accuracy. As well as, the AGender-Tune system improved the classification accuracy by approximately 3% compared with best reported results of the state-of-the-art.

Table 16. Overall performance comparison in speaker's age and gender classification.

	System	**Overall Acc. (%)**
[85]	(1) GMM base	43.1
	(2) Mean Super Vector	42.6
	(3) MLLR Super Vector	36.2
	(4) TPP Super Vector	37.8
	(5) SVM Base	44.6
	MFuse 1+2	45.2
	MFuse 3+4	40.3
	MFuse 1+2+3+4	50.4
	MFuse 1+2+3+4+5	52.7
[88]	(6) MFCCs-GMM	42.4
	(7) PLP-GMM Perceptual Linear Prediction	41.2

	(8) Temporal Patterns TRAPS-GMM	39.4
	(9) Prosodic, Voiced and Unvoiced segments, pitch period, jitter, shimmer, pauses, duration Total 219 features	40.6
	(10) Glottal Excitation, Harshness, hoarsness, increased strain, higher incidense of voice breaks, vocal tremor.	37.3
	Early fusion (7+8+9+10)	45.9
	Late fusion (7+8+9+10)	48.9
[89]	Age only, aGender and other DB for training Late fusion of acoustic and prosodic	51.2
T-MFCCs class models based	**T-MFCCs-I-vector**	**56.13**
	(6) T-MFCCs-DNN	**58.98**
Speakers Models	**(7) SDC- Class Model**	**47.7**
	(8) SDC-Speaker Model	**49.3**
	Our fused model (7+8)	**57.21**
AGender-Tune System class models based	**AGender-Tune**	**55.16**
AGender-Tune System speaker models based	**AGender-Tune**	**57.55**
T-MFCCs speaker models based	**(9) T-MFCCs-DNN**	**59.59**
Fusion	**Our fused model (6+9)**	**61.16**

The T-MFCCs set is proved to be more effective than the traditional MFCCs features in speaker's age and gender classification. There are two main reasons behind this improvement. First, introducing phoneme labels to create T-MFCCs for age and gender problem has a significant impact on the MFCCs, which become more discriminative and descriptive. By phoneme labels, phonetic components in a speaker's speech signal have been captured and used in detecting the speaker's age and gender information. Second, the regularized weights converged faster and provided higher variance between classes. These improvements boosted the performance of the classifiers.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

The goal of this research is to improve the classification accuracies in speaker's age and gender classification. For this purpose, major contributions are made to the area of feature extraction and classifier design. First, a novel approach is introduced to generate T-MFCCs feature set by using DNNs. The proposed system uses HTK to find tied-state triphones for all utterances, which are used as labels for the output layer in the DNNs for the first time in age and gender classification. The tied-state triphone labels are obtained through the forced alignment of trained GMM based on hidden Markov models (HMMs) by using both maximum likelihood (ML) and minimum phone error (MPE) techniques. The phoneme labels are used to capture the phonetic components in the speech. The involvement of the phoneme labels in the generation of the transformed MFCCs made it possible to grasp the prosodic features, such as intonation, stress, tone, and rhythm, of a speaker. As well as, the transformed features are the result of using phoneme labels in the training data, and this helped to remove any noise or silent frames so that the transformed features are calculated without acoustic background noise. To improve the performance of the traditional MFCCs, the transformed MFCCs feature set is generated by using BNF extractor. In the BNF extractor, phoneme labels are used to capture phonetic components in the speech. We showed that the DNN can be designed and trained to adapt smoothly with the BNF extractor, so that new transformed features can be obtained. The employment of the BNF has several benefits as eliminating the redundant values from the input feature set by reducing the number of units inside the bottleneck layer and reflecting the class labels during the classification process. Moreover, the bottleneck layer forces the neural

layers to filter the input features to keep the descriptive and distinctive features derived from short speech utterances. The adult classes represent a wide range of ages between 25 and 54 years old for female and male speakers. While the lower end of the adult classes is close to youth classes, the upper end is close to senior classes. The T-MFCCs features have good variance among female classes. It is reflected as an increase in the classification accuracies for female speakers. On the other hand, the MFCCs and T-MFCCs features have less variance among the male classes compared to that of female classes. As a result, misclassification occurred among male classes, especially between young and adult male classes. Introduction of shared class labels among misclassified classes to regularize the weights in DNN, the shared labels are proposed to regularize weights between DNN layers. The regularized weights provided faster convergence and higher variance between classes. The performance evaluation of the new features is done by several classifiers such as, DNN, GMM-UBM, and I-Vector. It is observed that the transformed MFCCs are more effective than the traditional MFCCs in speaker's age and gender classification.

Second, the DNN-based speaker models using the SDC feature set was proposed in order to improve the classification accuracies. The proposed method creates a model for each speaker in the training set. In the testing phase, for each speech utterance, the similarity between the test utterance model and the speaker class models are compared. Two feature sets have been used: Mel-frequency cepstral coefficients (MFCCs) and shifted delta cepstral (SDC) coefficients. This work aims to build a model for each speaker instead of using one model for each class of speakers, whom belong to the same class of age and gender. Speakers models approach have proved its ability to capture the unique characteristics of the speaker more efficiently than creating a model that consists of a set

of speakers. The possible benefit of a speaker-based model is that the system can use all the features of speakers who belong to the same class to improve the classification accuracies. The proposed model by using the SDC feature set achieved better classification results than that of MFCCs. The experimental results showed that the proposed SDC speaker model + SDC class model outperformed all the other systems. The proposed speaker models show a better performance while classifying challenging middle-aged female and male classes where the other methods fail to classify.

Third, AGender-Tune system was proposed by fine-tuning two DNN architectures. The first DNN is the Age DNN which is used to classify four groups of age, the second DNN is the Gender DNN which is used to classify the gender. Then, the two pre-trained DNNs are reused to tune a third DNN which can classify the age and gender of the speaker together. The input for the third DNN is the element-wise summation of the output layers of the Age and the Gender DNNs. The results of the proposed work are compared with two baseline systems; the I-Vector and GMM-UBM on a public database.

To evaluate the performance of the proposed methods for speaker age and gender classification, several experiments have been conducted on a public database. Experimental results show the effectiveness of the proposed methods in enhancing the classification accuracy.

REFERENCES

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 28, pp. 357-366, 1980.

[2] H.-J. Kim, K. Bae, and H.-S. Yoon, "Age and gender classification for a home-robot service," *The 16th IEEE International Symposium on Robot and Human interactive Communication*, 2007, pp. 122-126.

[3] D. Matrouf, N. Scheffer, B. G. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *INTERSPEECH*, 2007, pp. 1242-1245.

[4] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech," in *Odyssey*, 2010, pp. 28-33.

[5] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *INTERSPEECH*, 2011, pp. 857-860.

[6] D. Reynolds, "Gaussian Mixture Models," in *Encyclopedia of Biometrics*, ed: Springer, 2009, pp. 659-663.

[7] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication,* vol. 12, pp. 247-251, 1993.

[8] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language,* vol. 14, pp. 373-400, 2000.

[9] S.-L. Wu, E. Kingsbury, N. Morgan, and S. Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing,* 1998, pp. 721-724.

[10] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Elsevier Encyclopedia of Language and Linguistics*, 2005.

[11] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, pp. 2067-2080, 2011.

[12] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *IEEE Spoken Language Technology Workshop (SLT),* 2012, pp. 366-369.

[13] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2014, pp. 2494-2498.

[14] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 4960-4964.

[15] S. Kundu, G. Mantena, Y. Qian, T. Tan, M. Delcroix, and K. C. Sim, "Joint acoustic factor learning for robust deep neural network based automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2016, pp. 5025-5029.

[16] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, and H. Ney, "A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2017, pp. 2462-2466.

[17] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2014, pp. 5337-5341.

[18] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," Proceedings of the IEEE, vol. 101, pp. 1136-1159, 2013.

[19] B. M. L. Srivastava, H. Vydana, A. K. Vuppala, and M. Shrivastava, "Significance of neural phonotactic models for large-scale spoken language identification," in *International Joint Conference on Neural Networks (IJCNN),* 2017, pp. 2144-2151.

[20] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language,* vol. 27, pp. 209-227, 2013.

[21] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "Shifted-delta mlp features for spoken language recognition," *IEEE Signal Processing Letters,* vol. 20, pp. 15-18, 2013.

[22] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP),* 2010, pp. 5270-5273.

[23] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 271-284, 2007.

[24] T. W. Buchanan, K. Lutz, S. Mirzazade, K. Specht, N. J. Shah, K. Zilles*, et al.*, "Recognition of emotional prosody and verbal components of spoken language: an fMRI study," *Cognitive Brain Research,* vol. 9, pp. 227-238, 2000.

[25] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'01)*, 2001, pp. 343-346.

[26] T. Wu, J. Duchateau, J.-P. Martens, and D. Van Compernolle, "Feature subset selection for improved native accent identification," *Speech Communication,* vol. 52, pp. 83-98, 2010.

[27] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies,* 2005, pp. 139-143.

[28] M. Piat, D. Fohr, and I. Illina, "Foreign accent identification based on prosodic parameters," in *Proceedings of INTERSPEECH*, 2008, pp. 759-762.

[29] J. Hou, Y. Liu, T. F. Zheng, J. Olsen, and J. Tian, "Multi-layered features with SVM for Chinese accent identification," in *International Conference on Audio Language and Image Processing (ICALIP),* 2010, pp. 25-30.

[30] C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary continuous speech recognition," *International Journal of Speech Technology,* vol. 7, pp. 141-153, 2004.

[31] A. Lazaridis, E. Khoury, J.-P. Goldman, M. Avanzi, S. Marcel, and P. N. Garner, "Swiss French regional accent identification," In *Proceedings of odyssey on the speaker and language recognition workshop,* 2014.

[32] S. Safavi, A. Hanani, M. Russell, P. Jancovic, and M. J. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification," *IEEE Signal Processing Letters,* vol. 19, pp. 829-832, 2012.

[33] A. DeMarco and S. J. Cox, "Iterative classification of regional British accents in i-vector space," in *Symposium on machine learning in speech and language processing*, 2012.

[34] K. Amino and T. Osanai, "Native vs. non-native accent identification using Japanese spoken telephone numbers," *Speech Communication,* vol. 56, pp. 70-81, 2014.

[35] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review,* vol. 43, pp. 155-177, 2015.

[36] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia,* vol. 16, pp. 2203-2213, 2014.

[37] D. Gharavian, M. Sheikhan, A. Nazerieh, and S. Garoucy, "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network," *Neural Computing and Applications,* vol. 21, pp. 2115-2126, 2012.

[38] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH*, pp. 223–227, 2014.

[39] M. Kotti and F. Paternò, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *International journal of speech technology,* vol. 15, pp. 131-150, 2012.

[40] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Transactions on Affective Computing,* vol. 6, pp. 69-75, 2015.

[41] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication,* vol. 57, pp. 1-12, 2014.

[42] S. Singh and E. Rajan, "Vector quantization approach for speaker recognition using MFCC and inverted MFCC," *International journal of computer applications,* vol. 17, pp. 0975-8887, 2011.

[43] M. K. Gill, R. Kaur, and J. Kaur, "Vector quantization based speaker identification," *International journal of computer applications,* vol. 4, pp. 0975-8887, 2010.

[44] F. K. Soong, A. E. Rosenberg, B. H. Juang, and L. R. Rabiner, "Report: A vector quantization approach to speaker recognition," *Bell Labs Technical Journal,* vol. 66, pp. 14-26, 1987.

[45] T. Kinnunen and P. Fränti, "Speaker discriminative weighting method for VQ-based speaker identification," in *Audio-and Video-Based Biometric Person Authentication*, 2001, pp. 150-156.

[46] J. He, L. Liu, and G. Palm, "A discriminative training algorithm for VQ-based speaker identification," *IEEE Transactions on speech and audio processing,* vol. 7, pp. 353-356, 1999.

[47] V. Radová and Z. Švenda, "Speaker identification based on vector quantization," in *Text, Speech and Dialogue*, 1999, pp. 83-83.

[48] H. Kekre, V. Bharadi, A. Sawant, O. Kadam, P. Lanke, and R. Lodhiya, "Speaker recognition using Vector Quantization by MFCC and KMCG clustering algorithm," in *International Conference on Communication, Information & Computing Technology (ICCICT)*, 2012, pp. 1-5.

[49] Z.-X. Yuan, B.-L. Xu, and C.-Z. Yu, "Binary quantization of feature vectors for robust text-independent speaker identification," *IEEE Transactions on Speech and Audio Processing,* vol. 7, pp. 70-78, 1999.

[50] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE transactions on Speech and Audio Processing,* vol. 3, pp. 72-83, 1995.

[51] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing,* vol. 10, pp. 19-41, 2000.

[52] P. Matejka and P. Schwarz, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 1979-1986, 2007.

[53] S. Fine, J. Navratil, and R. A. Gopinath, "A hybrid GMM/SVM approach to speaker identification," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, 2001, pp. 417-420.

[54] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM)," in *A Speaker Odyssey-The Speaker Recognition Workshop*, 2001, pp. 52–55.

[55] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America,* vol. 123, p. 3878, 2008.

[56] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 1435-1447, 2007.

[57] P. Kenny, T. Stafylakis, J. Alam, P. Ouellet, and M. Kockmann, "Joint factor analysis for text-dependent speaker verification," in *Proceedings Odyssey Workshop*, 2014, pp. 1-8.

[58] C. Yu, G. Liu, S. Hahm, and J. H. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014,* pp. 4017-4021.

[59] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014,* pp. 225-229.

[60] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing,* vol. 20, pp. 30-42, 2012.

[61] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*, pp. 1695-1699.

[62] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," *arXiv preprint arXiv:1504.00923,* 2015.

[63] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*, pp. 4052-4056.

[57] P. Kenny, T. Stafylakis, J. Alam, P. Ouellet, and M. Kockmann, "Joint factor analysis for text-dependent speaker verification," in *Proc. Odyssey Workshop*, 2014, pp. 1-8.

[58] C. Yu, G. Liu, S. Hahm, and J. H. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 4017-4021.

[59] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 225-229.

[60] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing,* vol. 20, pp. 30-42, 2012.

[61] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 1695-1699.

[62] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," *arXiv preprint arXiv:1504.00923,* 2015.

[63] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 4052-4056.

[64] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2015, pp. 4814-4818.

[65] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, 2009, pp. 1096-1104.

[66] M. McLaren, Y. Lei, N. Scheffer, and L. Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions," in *Proceedings of INTERSPEECH*, 2014, pp. 686-690.

[67] M. Black, A. Katsamanis, C.-C. Lee, A. C. Lammert, B. R. Baucom, A. Christensen, et al., "Automatic classification of married couples' behavior using audio features," in *INTERSPEECH*, 2010, pp. 2030-2033.

[68] P. Nguyen, D. Tran, X. Huang, and D. Sharma, "Automatic speech-based classification of gender, age and accent," in *Pacific Rim Knowledge Acquisition Workshop*, 2010, pp. 288-299.

[69] T. Schultz, "Speaker Characteristics," in Speaker Classification I: Fundamentals, Features, and Methods, C. Müller, *Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg*, 2007, pp. 47-74.

[70] M. Ranzato and G. E. Hinton, "Modeling pixel means and covariances using factorized third-order Boltzmann machines," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2551-2558.

[71] C. Ekanadham, S. Reader, and H. Lee, "Sparse deep belief net models for visual area V2," *Advances in Neural Information Processing Systems*, vol. 20, pp. 873-880, 2008.

[72] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 30-42, 2012.

[73] T. Deselaers, S. Hasan, O. Bender, and H. Ney, "A deep learning approach to machine transliteration," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 2009, pp. 233-241.

[74] D. Yu, S. Wang, Z. Karam, and L. Deng, "Language recognition using deep-structured conditional random fields," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5030-5033.

[75] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, pp. 1527-1554, 2006.

[76] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, pp. 1-127, 2009.

[77] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, et al., "Developments and directions in speech recognition and understanding, Part 1 [DSP Education]," *IEEE Signal Processing Magazine*,vol. 26, pp. 75-80, 2009.

[78] E. D. Mysak, "Pitch and duration characteristics of older males," *Journal of Speech & Hearing Research*, 1959.

[79] N. Minematsu, M. Sekiguchi, and K. Hirose, "Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2002, pp. I-137-I-140.

[80] C. Muller, F. Wittig, and J. Baus, "Exploiting speech for recognizing elderly users to respond to their special needs," in *Eighth European Conference on Speech Communication and Technology*, 2003, pp. 1305-1308.

[81] W. Spiegl, G. Stemmer, E. Lasarcyk, V. Kolhatkar, A. Cassidy, B. Potard, et al., "Analyzing features for automatic age estimation on cross-sectional data," In *INTERSPEECH* 2009, pp. 2923-2926.

[82] Ajmera, J., Burkhardt, F., 2008. Age and gender classification using modulation cepstrum. In: Proc. *Odyssey*, p. 025

[83] C. A. Müller and F. Burkhardt, "Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age," In *INTERSPEECH*, 2007, pp. 2277-2280.

[84] M. K. Wolters, R. Vipperla, and S. Renals, "Age recognition for spoken dialogue systems: Do we need it?," in *INTERSPEECH*, 2009, pp. 1435-1438.

[85] Li, M., Han, K. J., & Narayanan, S. "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27(1), pp.151-167, 2013.

[86] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, et al., "Comparison of four approaches to age and gender recognition for telephone applications," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. 1089-1092

[87] M.-W. Lee and K.-C. Kwak, "Performance comparison of gender and age group recognition for human-robot interaction," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 3(12), pp.207-211, 2012.

[88] T. Bocklet, G. Stemmer, V. Zeissler, and E. Nöth, "Age and gender recognition based on multiple systems-early vs. late fusion," in *INTERSPEECH*, 2010, pp. 2830-2833.

[89] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features," in *INTERSPEECH*, 2010, pp. 2818-2821.

[90] M. H. Bahari and H. Van Hamme, "Speaker age estimation and gender detection based on supervised non-negative matrix factorization," in *IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, 2011, pp. 1-6.

[91] R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano, "Public speech-oriented guidance system with adult and child discrimination capability," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, 2004, pp. I-433.

[92] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1975-1985, 2011.

[93] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504-507, 2006.

[94] L. R. Rabiner and R. W. Schafer, “Digital Processing of Speech Signals,” *Prentice-Hall*, pp 375-377, 1978.

[95] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, et al., "The HTK book," *Cambridge University Engineering Department, HTK version 3.4 edition*, December 2006.

[96] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, pp. 1771-1800, 2002.

[97] A.-r. Mohamed, D. Yu, and L. Deng, "Investigation of full-sequence training of deep belief networks for speech recognition," in *INTERSPEECH*, 2010, pp. 2846-2849.

[98] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *IEEE Transactions on Signal and Information Processing*, 3:e2, 2014.

[99] A. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5060-5063.

[100] Y. Bao, H. Jiang, C. Liu, Y. Hu, and L. Dai, "Investigation on dimensionality reduction of concatenated features with deep neural network for LVCSR systems," in *IEEE 11th International Conference on Signal Processing (ICSP)*, 2012, pp. 562-566.

[101] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. 757-760.

[102] F. Grézl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4729-4732.

[103] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, et al., "The INTERSPEECH 2010 paralinguistic challenge," in *INTERSPEECH*, 2010, pp. 2795-2798.

[104] Burkhardt F, Eckert M, Johannsen W, Stegmann J A Database of Age and Gender Annotated Telephone Speech. in *Proceeding of 7th International Conference on Language Resources and Evaluation (LREC)*, 2010, pp 1562-1565

[105] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine Learning*, vol. 45, pp. 171-186, 2001.

[106] B. D. Barkana and J. Zhou, "A new pitch-range based feature set for a speaker's age and gender classification," *Applied Acoustics*, vol. 98, pp. 52-61, 2015.